Daniele Sereno, Pietro Valocchi

CODIFICA NUMERICA DEL SEGNALE AUDIO

CODIFICA NUMERICA
DEL SEGNALE AUDIO

Daniele Sereno, Pietro Valocchi

CODIFICA NUMERICA DEL SEGNALE AUDIO

PRESENTAZIONE

L'uomo persegue da sempre l'obiettivo di stabilire interrelazioni globali superando, con l'impiego di sofisticate tecnologie di trasmissione e di memorizzazione, le naturali barriere di distanza e di tempo. Questo sogno sta rapidamente divenendo realtà con lo sviluppo di nuove applicazioni multimediali interattive attraverso reti di telecomunicazione a copertura planetaria. Nell'ambito delle comunicazioni multimediali, l'audio costituisce il medium che più contribuisce a creare l'effetto presenza nell'interazione a distanza e, negli anni, le comunicazioni vocali hanno certamente rappresentato il principale servizio offerto dai sistemi di telecomunicazione. L'evoluzione tecnologica ha ovviamente avuto un profondo impatto sulla codifica di tali segnali, aprendo nuovi ed interessanti scenari applicativi con il passaggio da sistemi di tipo analogico a codifiche di tipo numerico, adatte al trattamento mediante elaboratore. In un intervallo di tempo relativamente breve, si è assistito al passaggio dalla trasmissione numerica su rete fissa ai sistemi di comunicazione wireless, alla realizzazione di reti integrate dati ed audio, all'introduzione di servizi audio numerici a larga banda sia diffusivi che interattivi e così via. Questa riconfermata importanza dell'audio nel mondo dell'Information and Communication Technology, unita alla constatazione della frammentarietà della documentazione preesistente, ha suggerito l'idea di sviluppare un testo organico sull'argomento, con l'intento di fornire, a quanti operano nel settore, un'aggiornata visione dello stato dell'arte e delle prospettive di sviluppo di prodotti e servizi, quale supporto di riferimento per le decisioni e per le esigenze legate all'operatività aziendale.

Rimandando al seguito per l'adeguata presentazione dei contenuti, mi preme qui sottolineare uno degli aspetti che più caratterizzano questo volume, che va a pieno titolo ad arricchire la collana editoriale della SSGRR. Nel libro sono infatti confluite le competenze di un centro di formazione post-universitaria, qual è appunto la Scuola Superiore G. Reiss Romoli, e le competenze di un centro di ricerche avanzate, qual è appunto

lo CSELT. È questo un significativo esempio di collaborazione tra formazione e ricerca al servizio delle aziende, con tutte le premesse, quindi, per contribuire efficacemente al processo continuo di aggiornamento di quanti operano nel competitivo contesto delle telecomunicazioni.

Agli autori va il sentito ringraziamento mio e della Scuola tutta per il grande impegno profuso nella realizzazione del testo al quale non può che augurarne il migliore successo di conoscenza e diffusione, tenuto conto della ricchezza del suo contenuto.

Antonio Zappi
Amministratore Delegato
Scuola Superiore G. Reiss Romoli

INDICE

APPENDICI

PREFAZIONE

Il linguaggio rappresenta ancora il mezzo di comunicazione più desiderabile per trasferire informazione tra esseri umani. Sebbene l'avvento della Information Technology ci abbia abituati e ci abituerà sempre più a considerare congiuntamente espressioni dell'informazione diverse, quali voce, immagini e testi, la comunicazione vocale riveste ancora un'importanza prioritaria. A seguito del salto tecnologico dal mondo analogico al mondo digitale, l'esigenza di comprimere più informazioni vocali su un generico canale di trasmissione numerica ha determinato lo sviluppo di tecniche di codifica vocale sempre più spinte, sfruttando in modo profondo le caratteristiche dell'apparato di produzione e di ricezione del segnale. Sebbene tale esigenza di alti fattori di compressione sia stata smussata dalla disponibilità di canali di trasmissione ad altissima capacità (fibre ottiche), lo sviluppo esponenziale dei sistemi di comunicazione mobile ha continuato a stimolare l'ideazione di tecniche sempre più efficienti.

In questo testo si vuole offrire un quadro per quanto possibile completo ed aggiornato delle tecniche di codifica numerica di segnali audio per la loro trasmissione o memorizzazione. In particolare, non ci si vuole limitare a fornire una descrizione, sia pur approfondita, delle tecniche esistenti, bensì evidenziare quei principi alla base della codifica audio digitale, per permettere al lettore di analizzare autonomamente anche tecniche non descritte nel dettaglio.

Nel seguito si fa riferimento genericamente a segnali audio, volendo comprendere in essi sia il segnale vocale (al quale la presentazione è principalmente rivolta) sia segnali audio a banda larga (es.: audio HiFi). Nel campo delle telecomunicazioni, questo è giustificato principalmente dallo sviluppo delle trasmissioni numeriche che mettono a disposizione dell'utente canali ad elevato bit-rate. In tal modo è possibile offrire servizi (es.: teleconferenza, Audio on Demand, ecc.) che richiedono segnali audio a banda maggiore di quella telefonica.

Inoltre, considerare oltre alle problematiche di trasmissione quelle di memorizzazione risulta una scelta naturale, visto che l'unica (anche se non trascurabile) differenza tra le due applicazioni è nel ritardo con il quale l'informazione scambiata viene utilizzata. Presentare oltre agli standard per

telecomunicazioni quelli per audio digitale dell'elettronica consumer (es.: Compact Disk, Digital Compact Cassette, ecc.), offre uno scenario più completo del settore della codifica per coprire probabili futuri sviluppi.

Il tema della codifica audio è affrontato per gradi di complessità crescente. Il primo capitolo è relativo all'analisi del segnale audio analogico. In particolare, per il segnale vocale, vengono presentate le caratteristiche sia della sorgente (l'apparato vocale) sia della destinazione (l'apparato uditivo). Come sarà mostrato nel seguito (cap. 5 e 6), l'analisi del segnale vocale ha un ruolo fondamentale per lo sviluppo delle codifiche predittive e per modelli. D'altra parte, le caratteristiche dell'udito sono alla base dei codificatori percettivi (cap. 7), in grado di raggiungere elevati livelli di compressione con un degrado impercettibile della qualità del suono. Il primo capitolo è completato da una rapida presentazione delle tecniche di trasduzione elettro-acustica e delle caratteristiche del canale telefonico tradizionale.

La conversione analogico-digitale viene trattata nel secondo capitolo, separatamente dalle tecniche di codifica. La trasformazione della sorgente da analogica a discreta è infatti il punto di partenza per eventuali successive tecniche di compressione. La trattazione ha il fine di evidenziare quale siano le alterazioni subite dal segnale per capire come sia possibile minimizzarle in fase di codifica. Il capitolo è completato con un accenno sugli aspetti implementativi dei convertitori A/D e D/A.

Eseguito il mappaggio della sorgente da analogica a discreta, nel terzo capitolo vengono richiamati alcuni risultati fondamentali della teoria dell'informazione. Lo scopo è innanzitutto quello di mostrare i limiti teorici sui rapporti di compressione ottenibili; inoltre, in tale ambito vengono presentate le tecniche per ridurre il flusso numerico di una sorgente discreta senza ridurre l'entropia dell'informazione da essa emessa. In tale capitolo si da un accenno anche alla quantizzazione vettoriale, in quanto, pur essendo associata a perdita di informazione, risulta una tecnica di compressione poco legata alla specifica sorgente, al contrario delle tecniche presentate nel seguito.

Il successivo quarto capitolo completa la presentazione dei codificatori che lavorano su campioni isolati del segnale sfruttando differenti mappaggi da continuo a discreto. La compressione che si ottiene in tal modo avviene esclusivamente a spese di una degradazione del segnale che, però, può essere ritenuta trascurabile secondo specifici criteri di ottimizzazione.

Ulteriori incrementi del grado di compressione si possono ottenere trattando non più campioni isolati, ma blocchi di essi, come mostrato nel quinto capitolo. Anche in questo caso, è necessario distinguere tra tecniche in grado di ridurre il flusso numerico della sorgente a spese di una degradazione del segnale e tecniche che, come la codifica predittiva, permettono di eliminare esclusivamente la ridondanza in esso contenuta. La codifica predittiva è esaminata con un certo dettaglio in quanto, oltre alla sua importanza intrinseca, è di base per le tecniche di codifica per modelli presentate nel capitolo successivo.

Il sesto capitolo è, infatti, relativo alle codifiche per modelli. Tali tecniche sono un punto di arrivo delle tecniche di codifica che lavorano nel dominio del tempo ed infatti in esse si ritrovano tracce di quasi tutte le tecniche precedentemente esposte.

L'ultimo capitolo è relativo alle codifiche nel dominio della frequenza. Seguendo la loro evoluzione, vengono dapprima analizzati i codificatori nel dominio della frequenza di forma d'onda, che tendono ad una fedele ricostruzione del segnale d'ingresso. Nel seguito vengono analizzati i codificatori percettivi che, sfruttando le caratteristiche dell'apparato uditivo, introducono concetti innovativi sui criteri da adottare per valutare la qualità del segnale elaborato.

Il testo è focalizzato sulla parte algoritmica della codifica: gli aspetti implementativi, salvo casi di particolare importanza, sono trattati a livello descrittivo o rimandati in appendice. Gli algoritmi scelti sono stati selezionati tra quelli effettivamente utilizzati negli standard esistenti o tra quelli che maggiormente agevolano la comprensione delle tecniche impiegate. La trattazione matematica svolta nel testo presuppone nozioni di analisi matematica, processi aleatori, teoria ed elaborazione numerica dei segnali forniti normalmente da corsi a livello universitario.

1

IL SEGNALE AUDIO

1.1 TRASMISSIONE DEL SEGNALE AUDIO

La trasmissione (in tempo reale o meno) del segnale audio è necessaria per lo sfruttamento dell'informazione ad esso associata a soggetti non presenti nel luogo dove è situata la sorgente. Il sistema utilizzato per tale trasferimento nel seguito viene indicato come *canale audio*. Nel caso in cui il segnale audio sia un segnale vocale, il sistema utilizzato per la sua trasmissione viene indicato come *canale telefonico*.

Il segnale audio consiste in variazioni di pressione in funzione del tempo e, per essere trasmesso tramite un sistema di comunicazione, richiede innanzitutto la sua trasformazione in segnale elettrico analogico (*trasduzione*). Per la trasmissione di tale segnale analogico, si può o meno ricorrere ad una sua trasformazione in un flusso numerico (*conversione Analogico/Digitale*). I vantaggi di una trasmissione numerica sono ben noti (maggiore robustezza agli errori di trasmissione, minore criticità delle apparecchiature, predisposizione all'elaborazione numerica, ecc.) come è ben noto lo svantaggio principale: una banda considerevolmente superiore a quella richiesta dal segnale analogico. A fronte di questo inconveniente è necessario individuare opportune tecniche di *compressione* a cui assoggettare il flusso emesso dalla sorgente.

È auspicabile che tutte le trasformazioni utilizzate (trasduzione, conversione A/D, compressione) siano tali che le inevitabili degradazioni dell'informazione scambiata risultino "non apprezzabili". Per far questo è necessario, che il canale audio abbia caratteristiche (essenzialmente banda e dinamica) migliori di quelle rilevabili dal destinatario dell'informazione trasmessa: l'ap-

parato uditivo. Inoltre, nel caso in cui il segnale scambiato abbia caratteristiche inferiori a quelle potenzialmente utilizzabili dalla destinazione, come nella telefonia, le specifiche sul canale possono essere, ovviamente, rese meno stringenti. Per meglio comprendere i vincoli sul canale, dunque, è necessario analizzare sia le caratteristiche del destinatario dell'informazione trasmessa (l'*apparato uditivo*). Inoltre, per quel sottoinsieme del segnale audio relativo ai segnali vocali, che costituiscono la grande maggioranza dei segnali considerati nei sistemi di comunicazione tradizionali, risulta essenziale conoscere le caratteristiche della sorgente del segnale stesso (l'*apparato vocale*).

1.2 CARATTERISTICHE DELL'APPARATO UDITIVO

La percezione del suono avviene tramite due fasi. La prima fase è relativa alla trasformazione da parte dell'orecchio del suono da variazioni di pressione in impulsi nervosi, mentre la seconda fase è relativa all'interpretazione dello stimolo nervoso da parte del cervello. La fase di interpretazione assume un'importanza relativa maggiore man mano che il suono diventa via via più degradato, cioè sottoposto ad alterazioni che ne alterano le caratteristiche rispetto all'originale (es.: distorsioni, limitazioni di banda, ecc.). Questo crescente ruolo dell'interpretazione si manifesta con un crescente affaticamento del ricevente, un indice della qualità della trasmissione [Bon91].

Per studiare le caratteristiche della traduzione del suono da parte dell'orecchio e comprenderne i meccanismi, è necessario innanzitutto analizzarne la struttura. L'orecchio può essere scomposto in tre parti distinte: l'orecchio esterno, l'orecchio medio e l'orecchio interno (fig. 1.1).

L'orecchio esterno è formato dal padiglione e dal condotto uditivo. La funzione principale del padiglione è relativa alla localizzazione della sorgente. La posizione dei padiglioni permette di ottenere una buona risoluzione nel semipiano orizzontale posto di fronte all'ascoltatore (cioè, posizione della sorgente a destra o a sinistra dell'ascoltatore stesso). A tal fine vengono sfruttate essenzialmente le differenze temporali e di livello dei suoni percepiti da ciascun orecchio. La testa dell'ascoltatore, infatti, risulta trasparente per suoni a frequenza inferiore al kHz. Considerando, quindi, segnali a frequenza inferiore a tale soglia e fissata la velocità del suono a circa 340 m/s, per una sorgente posta sulla linea congiungente le due orecchie, la distanza tra i

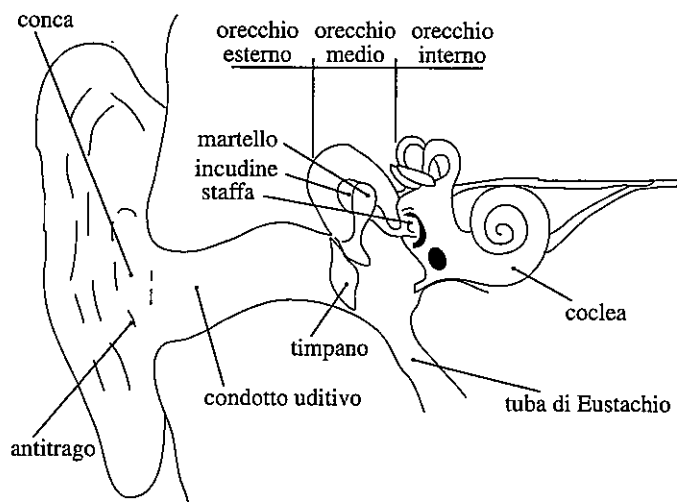



Fig. 1.1 - Struttura dell'apparato uditivo.

padiglioni si traduce in un ritardo di circa 500 μs tra i segnali da esse rilevati. Tale ritardo si manifesta come una differenza di fase tra il segnale rilevato dall'orecchio destro rispetto a quello rilevato dal sinistro, permettendo la localizzazione della sorgente. Per i segnali a frequenza maggiore di un kHz, invece, la testa dell'ascoltatore costituisce un ostacolo, che provoca un'attenuazione del segnale rilevato dall'orecchio sul lato opposto a quello dove è posizionata la sorgente. La localizzazione della sorgente, quindi, è possibile sfruttando la differenza di livello tra i segnali destro e sinistro.

Le differenze di fase e di livello tra i segnali destro e sinistro non permettono di localizzare una sorgente sul piano verticale passante di fronte, al di sopra e dietro l'ascoltatore (piano mediano). A tal fine vengono, invece, sfruttate le riflessioni prodotte dai rilievi presenti sul padiglione dell'ascoltatore. Tali riflessioni, che sono generate da parte della conca per i suoni frontali e dall'antitrago per quelli che provengono dall'alto, risultano in opposizione di fase con il suono incidente. Ciò introduce degli zeri nella funzione di trasferimento dell'orecchio esterno a frequenze dipendenti dall'entità del ritardo del suono riflesso su quello incidente. Ad esempio, nel caso di due sorgenti, una frontale ed una perfettamente perpendicolare all'ascoltatore, date le distanze tra l'ingresso del condotto uditivo e le pareti della conca e

dell'antitrago di circa 13 e 6 mm rispettivamente, il ritardo τ (relativo a percorsi di 26 e 12 mm) è rispettivamente di circa 75 e 35 μs. Gli zeri della funzione di trasferimento che cadono nella banda audio (posti a multipli dispari di $1/2\tau$) sono dunque a circa 7 e 20 kHz per la sorgente frontale e a 14 kHz per la sorgente perpendicolare all'ascoltatore. La localizzazione della sorgente sul piano mediano può, quindi, avvenire tramite lo spettro del segnale e sfruttando la presenza di tali zeri nella funzione di trasferimento.

Il secondo componente dell'orecchio esterno, il condotto uditivo, è un canale, approssimativamente circolare, aperto dalla parte del padiglione e chiuso all'altro estremo dal timpano. Nell'adulto la sua larghezza è di circa 0.7 cm, con una lunghezza di circa 2.7 cm. Un terzo del canale uditivo (quello esterno) è di natura cartilaginea, mentre i due terzi interni sono ossei. Il punto a sezione minore è nella congiunzione tra parte cartilaginea ed ossea (istmo) per cui la sua sezione è, in realtà, a forma di due coni riuniti per il loro apice tronco. Il condotto permette alle onde di pressione di raggiungere gli organi interni dell'orecchio con una funzione di trasferimento che, però, non risulta piatta in frequenza. Infatti, in un condotto chiuso ad una estremità ed aperto dall'altra, si instaurano risonanze che interessano segnali aventi lunghezza d'onda multipla di quattro volte la lunghezza del canale. Dalle dimensioni del condotto si ricava che la sua prima risonanza si ha per segnali a frequenza di circa 3 kHz. In conseguenza, per segnali a tale frequenza si ha un guadagno acustico tra ingresso del canale e timpano (e quindi un massimo di sensibilità) che può raggiungere i 10 dB.

Il condotto uditivo termina sul timpano, che è l'organo che separa l'orecchio esterno dall'orecchio medio. L'orecchio medio è costituito da una cavità di circa 2 cm^3 (cassa del timpano) che racchiude la catena degli ossicini (martello, incudine e staffa). La cassa del timpano (contenente aria) è limitata da un lato dal timpano e dall'altro dalle strutture dell'orecchio interno. La pressione al suo interno è regolata dalla tuba di Eustachio, che la collega alla gola. Il timpano è responsabile della trasduzione del suono da variazioni di pressione a lavoro meccanico. Esso è costituito da una membrana cartilaginea, leggermente ellittica, di circa 9 mm di diametro. Le vibrazioni prodotte sul timpano dalle onde di pressione incidenti sono trasmesse dagli ossicini verso l'orecchio interno. Tramite un effetto leva da parte degli ossicini, ma soprattutto grazie alla differenza di superfici tra timpano e punto di appoggio

sulla coclea (finestra ovale), la trasmissione meccanica avviene con un guadagno in pressione di circa 15 volte.

Anche la funzione di trasferimento degli ossicini non è piatta, ma presenta una risonanza nell'intorno del kHz. Inoltre, gli ossicini sono tra loro vincolati tramite piccoli muscoli. Il primo di questi (tensore del timpano) è collegato al martello ed ha essenzialmente la funzione di tendere il timpano, al quale il martello stesso è connesso. Il secondo legamento (stapedio) è collegato alla giunzione tra incudine e staffa, regolando la sensibilità dell'orecchio. Infatti, al fine di evitare danni dovuti a sovra pressioni, lo stapedio irrigidisce tale giunzione in presenza di eccitazioni di livello superiore di 85 dB rispetto al livello minimo udibile. Questo meccanismo ha tempi di intervento che sono tra 50 e 150 ms e tempi di rilassamento di alcuni secondi. Esso, quindi, pur riducendo la sensibilità dell'udito in presenza di segnali a livelli elevati, non è purtroppo in grado di evitare i danni che possono essere prodotti da suoni impulsivi, anche se di livello modesto.

La coclea costituisce l'orecchio interno ed è responsabile della generazione degli stimoli nervosi da parte dell'orecchio verso il cervello. Essa è composta da un canale spiraliforme lungo circa 35 mm e composto longitudinalmente da tre cavità: la scala vestibolare, la scala timpanica e la scala media. La scala vestibolare e scala timpanica, se sviluppate linearmente, hanno entrambe una struttura approssimativamente conica (fig. 1.2), con una sezione iniziale (dal lato dell'orecchio medio) di 4 mm^2 e finale di 1 mm^2. Queste due cavità sono comunicanti nella parte a sezione minore (elicotrema), mentre nell'estremità a sezione maggiore sono terminate da due membrane: la scala vestibolare termina con la finestra ovale (sulla quale si appoggia la staffa), mentre la scala timpanica termina con la finestra tonda. In esse è contenuto un liquido viscoso (perilinfa) che fa si che le vibrazioni trasmesse dalla finestra ovale le attraversino interamente fino a provocare moti complementari sulla finestra tonda.

La scala media è responsabile della rilevazione dei suoni. Essa è formata da una cavità a sezione crescente man mano che ci si allontana dall'orecchio medio e contenente, anch'essa, un liquido (endolinfa). La scala media si appoggia alla scala vestibolare e a quella timpanica tramite due membrane, che sono rispettivamente la membrana di Reissner e la membrana basilare. Sulla membrana basilare è disposto l'organo del Corti, responsabile della genera-

zione degli stimoli nervosi. L'organo del Corti è costituito da circa 30000 organi sensibili (celle cigliate) disposte longitudinalmente lungo 4 file (fig. 1.2). Una di queste file (celle cigliate interne) è posizionata all'interno del vertice formato dalla congiunzione delle tre scale e ad essa è collegata la maggioranza delle terminazioni del nervo acustico (sinapsi). Le rimanenti file (celle cigliate esterne) sono raggruppate in una posizione più centrale della membrana basilare. Al di sopra delle celle cigliate vi è un'ulteriore membrana gelatinosa (membrana tettoriale), interna alla scala media, che è solidale alla struttura ossea della coclea e che si estende a coprire le celle cigliate stesse.

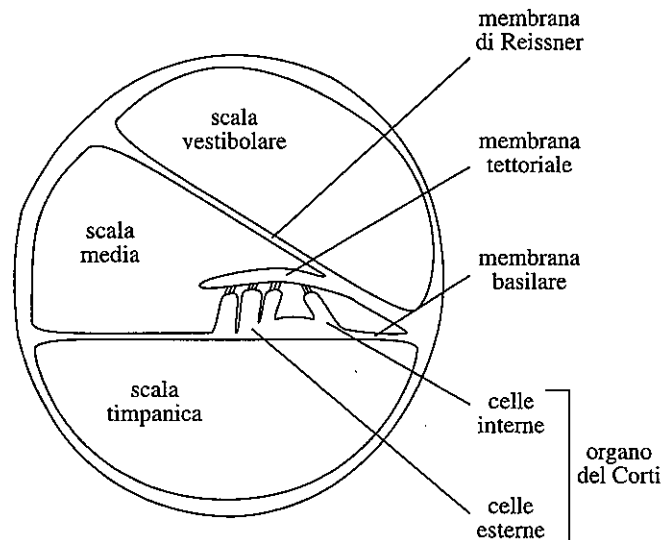



Fig. 1.2 - Struttura della coclea.

Il meccanismo di rilevazione dei suoni si basa sulle risonanze della membrana basilare provocate dai flussi interni alle scale vestibolari e timpaniche. La membrana basilare è più rigida e sottile verso il timpano e più spessa verso l'estremità apicale. Le sue caratteristiche di risonanza, dunque, variano in funzione della distanza dalla staffa. Analizzando il comportamento della membrana basilare a segnali sinusoidali puri isolati (toni), si nota come questi provochino delle risonanze con un inviluppo che cresce progressivamente in ampiezza allontanandosi dall'orecchio medio e, dopo aver raggiunto

un massimo, diminuisce bruscamente man mano che si prosegue verso l'elicotrema. La posizione del massimo sulla membrana basilare dipende dalla frequenza del segnale, allontanandosi dalla staffa al diminuire della frequenza del segnale (fig. 1.3). Tali oscillazioni (fortemente smorzate) fanno flettere le celle cigliate disposte tra la membrana basilare (mobile) e la membrana tettoriale (fissa). La compressione subita dalle ciglia aumenta la loro conducibilità, con una conseguente diminuzione della loro tensione interna. Tale variazione di potenziale provoca un'attivazione delle sinapsi del nervo acustico che producono stimoli nervosi verso il cervello. Le scariche di potenziale derivanti dall'attività delle sinapsi si trasformano in tal caso, da un'attività spontanea inferiore alle 100 scariche al secondo, in treni di scariche in corrispondenza degli istanti di maggiore compressione delle celle.

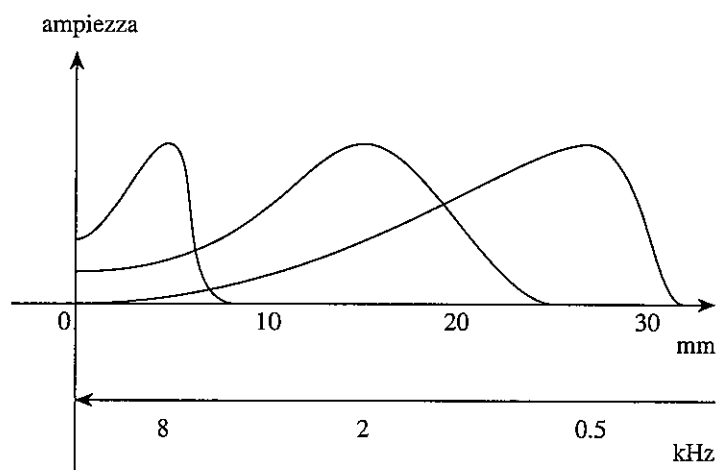



Fig. 1.3 - Inviluppo delle risonanze indotte sulla membrana basilare da toni a differente frequenza.

Il suono è, in definitiva, rilevato nelle sue componenti di ampiezza e frequenza tramite l'intensità degli impulsi nervosi emessi dalle celle cigliate e tramite la posizione nell'organo del Corti delle celle attive. L'attività delle celle, infatti, è funzione della posizione e dell'ampiezza e delle vibrazioni indotte sulla membrana basilare che, a loro volta, misurano la frequenza ed il livello delle componenti armoniche del suono.

Descritto il meccanismo di rilevazione dei suoni, risulta più agevole dare una spiegazione intuitiva alle prestazioni dell'udito sia in termini di sensibilità che di selettività. La caratterizzazione della sensibilità dell'udito al livello del segnale avviene tramite *diagrammi audiometrici* (fig. 1.4). Tali diagrammi sono costituiti da famiglie di curve isofoniche. Esse esprimono l'andamento della potenza di un tono di prova, al variare della sua frequenza, affinché l'intensità sonora percepita si mantenga costante e pari a quella di un tono di 1 kHz di livello fissato. Un incremento nella curva isofonica, quindi, è indice di una minore capacità uditiva e viceversa per una sua riduzione. La scala adottata per tali curve è logaritmica (dB) in quanto l'intensità sonora percepita cresce in proporzione logaritmica con l'intensità fisica del suono: una variazione di 3 dB della potenza del segnale porta, in media, a variazioni di intensità appena percepibili, mentre una variazione di 10 dB genera una sensazione di raddoppio del livello del suono.

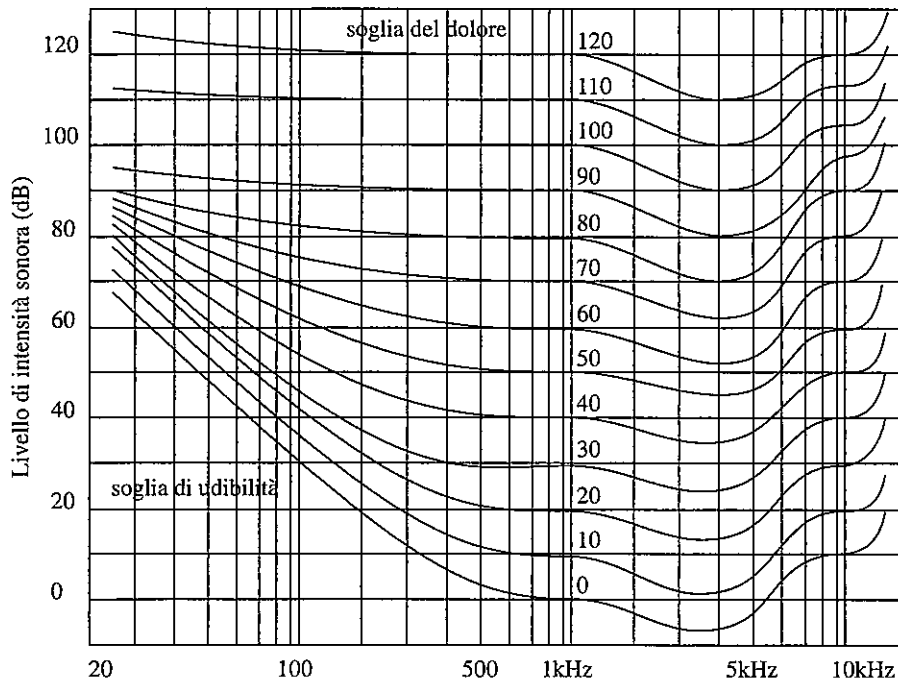



Fig. 1.4 - Audiogramma.

Nell'analisi di tali curve è necessario premettere che esse risultano molto variabili da soggetto a soggetto e, nello stesso soggetto, variano con le condizioni di affaticamento e negli anni. Analizzando i diagrammi audiometrici, comunque, si osserva che la sensibilità dell'udito al livello del segnale non è costante, ma dipende sia dalla frequenza che dal livello stesso. Per segnali di potenza elevata, le curve isofoniche sono relativamente piatte nella banda da 20 a 20000 Hz. Al diminuire del livello del segnale, le curve isofoniche accentuano la presenza di un massimo di sensibilità nell'intorno dei 3 kHz, dovuto alle risonanze presenti nell'orecchio medio ed esterno. Il più basso livello sonoro udibile viene indicato come *soglia di udibilità statica.* Tale livello di riferimento, posto a 0 dB, è standardizzato dall'ISO ad una pressione di 106^{-12} $6W/m^2$ per un segnale di frequenza pari ad un kHz. Il massimo livello utile viene indicato come "soglia del dolore." Con suoni di livello maggiore di quest'ultimo si rischiano danni irreversibili all'orecchio. Il valore della soglia del dolore è approssimativamente di 120 dB, corrispondente ad una potenza di circa 1 W. Per avere un'idea della potenza di tale segnale, si pensi che una normale conversazione tra due persone produce un segnale di circa 60 dB. Se si pensa che l'orecchio percepisce i segnali di livello prossimo alla soglia di udibilità con spostamenti del martello di frazioni di angstrom (10-8 cm), paragonabili al diametro dell'atomo di idrogeno, si ha anche un'idea di quanto sofisticato sia il meccanismo di trasduzione dell'apparato uditivo.

Altra caratteristica importante, oltre alla sensibilità dell'apparato uditivo al livello del segnale, è la sua risoluzione in frequenza o *selettività.* Essa è definita come la capacità di scomporre un segnale complesso nelle sue componenti spettrali. Per comprendere i meccanismi legati alla selettività può essere utile fare un parallelo tra il funzionamento dell'apparato uditivo con quello di uno strumento di misura. Come si è visto, la rilevazione delle componenti spettrali del segnale è legata alla rilevazione della posizione dei massimi di risonanza della membrana basilare. Se si introduce per la coclea il concetto di precisione nella misura della posizione di tali massimi, dovrebbe risultare evidente come non sia possibile rilevate variazioni di frequenza inferiori alla precisione dello strumento stesso. In pratica, l'apparato uditivo non è in grado di rilevare variazioni sulla posizione di un massimo di risonanza all'interno di un intervallo la cui ampiezza è dell'ordine del millimetro.

Per valutare come tale indeterminazione si traduca in termini di frequenze, è necessario innanzitutto osservare che la posizione "x" del massimo di risonanza sulla membrana non varia linearmente con la frequenza "f". Approssimativamente il loro legame è esprimibile come [Kin82]

$$f = 2.5 \cdot 10^{4-0.75x} \tag{1.1}$$

Per tale non linearità del mappaggio tra ascissa e frequenza, l'indeterminazione sulla posizione dei massimi di risonanza in un intervallo di ampiezza costante al variare della frequenza, si traduce in un'indeterminazione sulla frequenza dei segnali all'interno di una banda di ampiezze crescenti all'aumentare della frequenza del segnale stesso. Approssimativamente, l'ampiezza di tali bande può essere posta pari al 20% della loro frequenza centrale. Noto che un'ottava corrisponde ad un intervallo di frequenze gli estremi del quale sono in rapporto 2:1, tali bande hanno una spaziatura pari a circa 1/3 d'ottava. Di conseguenza, la risoluzione dell'apparato uditivo non è costante in frequenza, ma si riduce all'aumentare di questa e ciò giustifica l'utilizzo di scale logaritmiche per essa in acustica. Infatti anche il legame tra l'altezza del suono percepito e la frequenza del segnale è non lineare, ma può essere approssimato come [Del93]

$$f_{mel} = \frac{1000}{\log 2}\left(1 + \frac{f_{Hz}}{1000}\right) \tag{1.2}$$

L'unità di misura della grandezza in uscita da tale relazione è il *mel*. Ulteriore parametro da tenere in conto nell'analisi della selettività in frequenza, è la sua dipendenza dal livello del segnale. Infatti, dato che l'ampiezza dell'intervallo di risonanza aumenta all'aumentare del livello del segnale, la risoluzione peggiora all'aumentare del livello stesso.

L'analisi della sensibilità al livello del segnale e della selettività in frequenza dell'apparato uditivo è particolarmente importante in quanto da essi dipendono l'esistenza di fenomeni di *mascheramento*. Il mascheramento consiste nell'impossibilità di percepire componenti del segnale in presenza di componenti di maggiore potenza posti (nel tempo ed in frequenza) in loro prossimità. Nel caso di segnali di elevata potenza che precedono temporalmente dei segnali deboli (all'interno di un intervallo di circa 15 ms) si parla

di *mascheramento in avanti.* Per segnali mascheranti che seguono (all'interno di un intervallo di circa 2 ms) segnali deboli, si parla di *mascheramento all'indietro.* Nel caso in cui le componenti sotto esame siano contemporaneamente presenti nel segnale, ma in zone spettrali differenti, si parla di *mascheramento simultaneo.* Il mascheramento simultaneo può essere pensato, al pari della selettività, come una manifestazione della precisione della coclea nel processo di trasduzione del segnale audio. È infatti evidente come non sia possibile risolvere come distinti dei picchi di risonanza relativi a due componenti spettrali se la loro distanza è inferiore alla selettività della coclea. Di conseguenza, essi vengono fusi in un unico massimo, sulla posizione del quale ha maggiore influenza la componente a potenza maggiore.

I primi tentativi fatti per un'analisi quantitativa del mascheramento ricorrevano ad una modellizzazione dell'apparato uditivo come banco di filtri passabanda ideali, detti *filtri uditivi.* La tecnica utilizzata per ricavare le bande di tali filtri si basa sull'analisi del mascheramento prodotto su di un tono isolato da parte di rumore bianco filtrato passa banda (fig. 1.5). L'ipotesi alla base del procedimento è che il mascheramento dipenda dal rapporto tra l'energia del

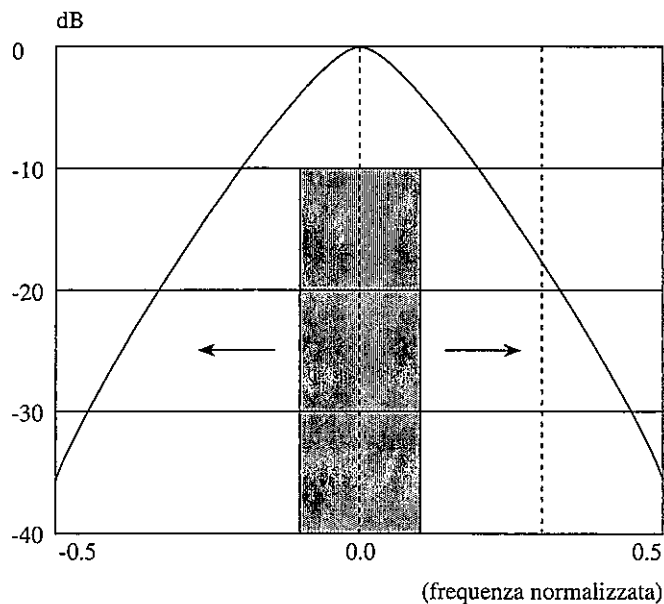



Fig. 1.5 - Costruzione della caratteristica di un filtro uditivo.

segnale mascherato e l'energia di quella parte dello spettro di potenza del segnale mascherante che cade all'interno dello stesso filtro uditivo nel quale è compreso il segnale. Per rilevare la caratteristica dei filtri uditivi, quindi, si può utilizzare un segnale sinusoidale e del rumore bianco con densità spettrale fissata. Allargando progressivamente la banda del rumore nell'intorno della frequenza del segnale mascherato, la potenza che contribuisce al mascheramento aumenta progressivamente, fino a che la banda del rumore stesso non supera la banda del filtro cercato. Superato tale valore l'efficacia del mascheramento non varia. Nella procedura indicata non risultano fissati i livelli del segnale mascherante e del mascherato. Per essi si fa l'ipotesi che il rapporto tra la potenza del segnale e quella del rumore che cade nella banda del filtro uditivo nel momento in cui il segnale mascherante ha il sopravvento sia costante e pari ad uno.

Le bande dei filtri uditivi, dette *bande critiche*, corrispondono ad intervalli della membrana basilare di ampiezza approssimativamente pari a 1.3 mm. Tale ampiezza è utilizzata come unità di misura dell'ascissa lungo la membrana basilare ed è definita *bark*. Le frequenze centrali "f" e le ampiezze "b" delle bande critiche sono riportate nella seguente tabella

bark	1	2	3	4	5	6	7	8	9	10	11	12
f(Hz)	50	150	250	350	450	570	700	840	1000	1170	1370	1600
b(Hz)	100	100	100	100	110	120	140	150	160	190	210	240

bark	13	14	15	16	17	18	19	20	21	22	23	24
f(Hz)	1850	2150	2500	2900	3400	4000	4800	5800	7000	8500	10500	13500
b(Hz)	280	320	380	450	550	700	900	1100	1300	1800	2500	3500

Tab. 1.1 - Frequenze centrali e ampiezze delle bande critiche.

In realtà, il modello dell'apparato uditivo presentato è notevolmente semplificato. Innanzitutto non esiste un unico banco di filtri centrati su frequenze prefissate, ma sarebbe necessario considerare per ciascun segnale un filtro centrato sulla sua frequenza. Inoltre i filtri uditivi non risultano né ideali, né lineari e possono esistere complesse interazioni tra componenti non appartenenti alla stessa banda. Per ulteriori approfondimenti su questi aspetti si rimanda alla bibliografia [Moo89].

Per migliorare il modello utilizzato per il mascheramento, è necessario definire la caratteristica di un generico filtro uditivo centrato su ciascuna componente spettrale del segnale. La caratteristica che se ne ricava [Moo89] ha un massimo abbastanza arrotondato, posizionato alla frequenza della componente spettrale e presenta un decadimento esponenziale a partire da esso. Una buona approssimazione della caratteristica si ottiene tramite la funzione

$$W(g) = (1 + p\,|g|)\,e^{-p|g|} \qquad (1.3)$$

che viene comunemente indicata come roexp(p) (rounded exponential). In tale espressione g rappresenta la frequenza normalizzata

$$g = \frac{(f - f_0)}{f_0} \qquad (1.4)$$

mentre p è un parametro che fissa la selettività del filtro (maggiore all'aumentare del valore di p); tali filtri risultano a banda crescente al crescere della frequenza centrale.

Si fa presente come anche tale modello sia semplificato. Infatti la caratteristica del filtro che si ottiene dalla roexp(p) è simmetrica, mentre quella dei filtri uditivi è generalmente asimmetrica, con un'attenuazione maggiore alle alte frequenze per i livelli inferiori dei segnali e viceversa per i livelli superiori. Inoltre, si hanno anche altre deviazioni della caratteristica reale dalla roexp(p) per le alterazioni introdotte dalla funzione di trasferimento dell'orecchio medio ed esterno ai livelli inferiori del segnale. Anche per questi aspetti si rimanda alla bibliografia citata [Moo89].

La rilevazione della caratteristica dei filtri uditivi è necessaria per valutare quantitativamente l'impatto che il mascheramento ha sulla sensibilità dell'apparato uditivo. Il mascheramento, infatti, si manifesta come un innalzamento del livello della soglia di udibilità nell'intorno della frequenza dei segnali mascheranti. La curva che si ottiene in tal modo è detta *soglia di udibilità dinamica* ed il contributo dei singoli segnali mascheranti è descritto da corrispondenti *caratteristiche di mascheramento* (fig. 1.6). Come le caratteristiche dei filtri uditivi siano legate alle caratteristiche di mascheramento può essere compreso facendo riferimento alla rilevazione sperimentalmente della soglia di udibilità dinamica.

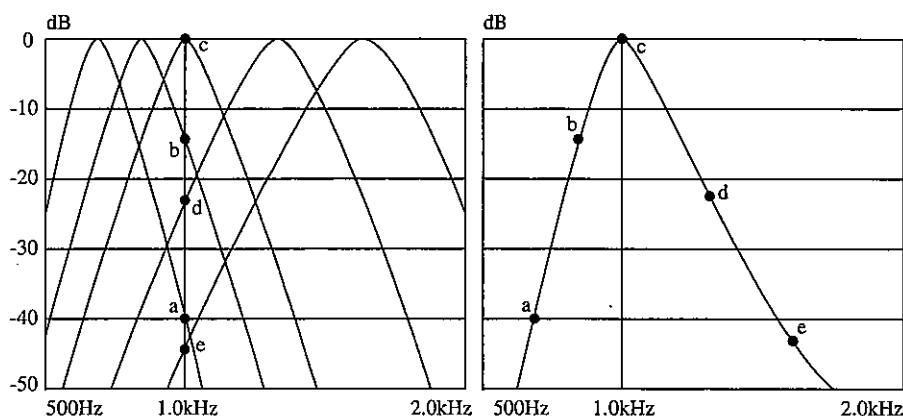



Fig. 1.6 - Costruzione di una caratteristica di mascheramento a partire da filtri uditivi.

Una possibile tecnica è quella di ricorrere ad un segnale mascherante a frequenza e livello fissati, registrando il livello minimo udibile di segnali di prova a frequenza variabile. Il segnale mascherante è solitamente rumore a banda limitata, mentre il segnale mascherato è un tono sinusoidale. È intuitivo pensare che il contributo al mascheramento dei segnali di prova da parte del mascherante sia proporzionale all'uscita dei relativi filtri uditivi. Fissata la potenza del segnale mascherante, l'entità del mascheramento per un segnale ad una data frequenza si può pensare proporzionale all'ampiezza che la caratteristica del relativo filtro uditivo assume alla frequenza centrale del segnale mascherante. L'andamento della caratteristica di mascheramento è quindi ricavabile per punti considerando una famiglia di filtri uditivi, ricavando per ciascuno di essi il livello alla frequenza del segnale mascherante e riportando tale livello alla frequenza centrale del filtro stesso [Moo89]. Dato che le caratteristiche dei filtri uditivi hanno banda crescente al crescere della frequenza centrale, le intersezioni risultano maggiori per le frequenze più elevate che non per le frequenze minori. Di conseguenza, la caratteristica di mascheramento risulta asimmetrica, con il ramo a frequenza maggiore che presenta una pendenza minore rispetto a quella del ramo a frequenza minore.

Se si considera come segnale mascherante un tono sinusoidale, le caratteristiche cambiano leggermente. In tale caso, infatti, si presentano battimenti che, con le loro variazioni periodiche di livello, possono rivelare la presenza

dei segnali mascherati. Le caratteristiche di mascheramento che si ottengono in questo caso, evidenziano il contributo dei battimenti con dei minimi in corrispondenza di multipli della frequenza del segnale mascherante (fig. 1.7).

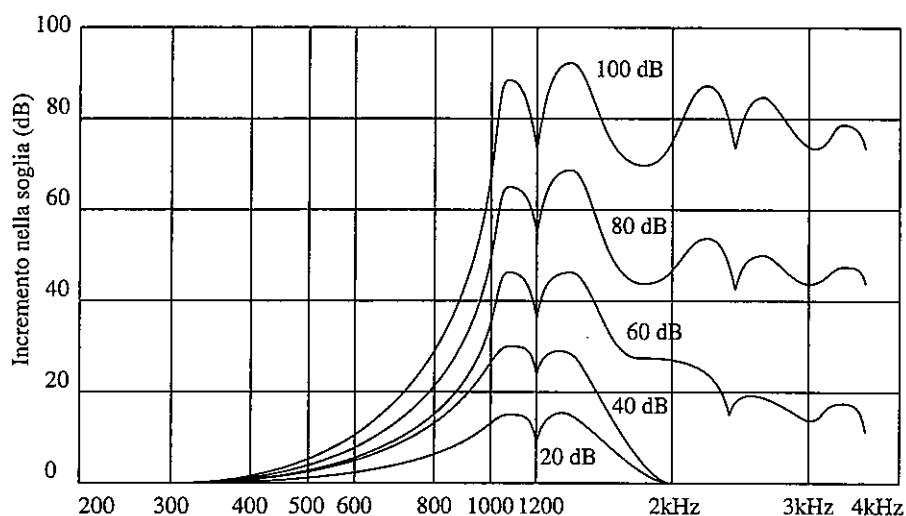



Fig. 1.7 - Effetto mascherante di un tono di 1200 Hz ad ampiezze differenti.

Ricavate le caratteristiche di mascheramento, la soglia di udibilità dinamica si ricava semplicemente come somma tra la soglia di udibilità statica e le caratteristiche dei vari segnali mascheranti (fig. 1.8).

Come mostrato nel seguito, l'innalzamento della soglia di udibilità dovuto al mascheramento è utilizzabile in codifica, tramite l'impiego di modelli percettivi, al fine di ridurre l'irrilevanza del segnale audio.

I fenomeni di mascheramento fin qui analizzati sono relativi a coppie di segnali. Nel caso di ascolto di un segnale audio in presenza di rumore ambientale bianco, invece, il mascheramento simultaneo precedentemente descritto si traduce in un innalzamento uniforme della soglia di udibilità di circa 20 dB al di sopra del livello del rumore stesso. Ciò è facilmente spiegabile se si pensa al contributo complessivo del rumore scomposto nel contributo delle sue componenti spettrali.

Oltre al mascheramento, è opportuno infine presentare un ulteriore fenomeno dell'apparato uditivo, che è quello della ricostruzione della

fondamentale. Questo fenomeno si manifesta nell'ascolto simultaneo di coppie di toni sinusoidali puri di frequenza differente, nel qual caso oltre ad essi si percepisce un terzo tono (fisicamente assente) la cui frequenza è tale da poter interpretare i due segnali generati come sue armoniche adiacenti. Ciò avviene a causa della non linearità della funzione di trasferimento dell'orecchio che provoca, durante la trasmissione del segnale verso la coclea, la generazione (per distorsione di intermodulazione) delle componenti con frequenze a combinazioni intere di quelle dei due segnali. Tale fenomeno risulta utile nei sistemi trasmissivi, dove i segnali vengono comunemente filtrati per eliminare la porzione inferiore della banda. Le componenti spettrali a frequenza inferiore, però, possono contenere informazioni essenziali dal punto di vista della qualità in quanto, ad esempio, legate alla fondamentale del segnale vocale. In tal caso, il fenomeno della ricostruzione della fondamentale permette di limitare la perdita di qualità derivante dal filtraggio passa alto, grazie alla ricostruzione della fondamentale stessa da parte delle sue armoniche a frequenze maggiore.

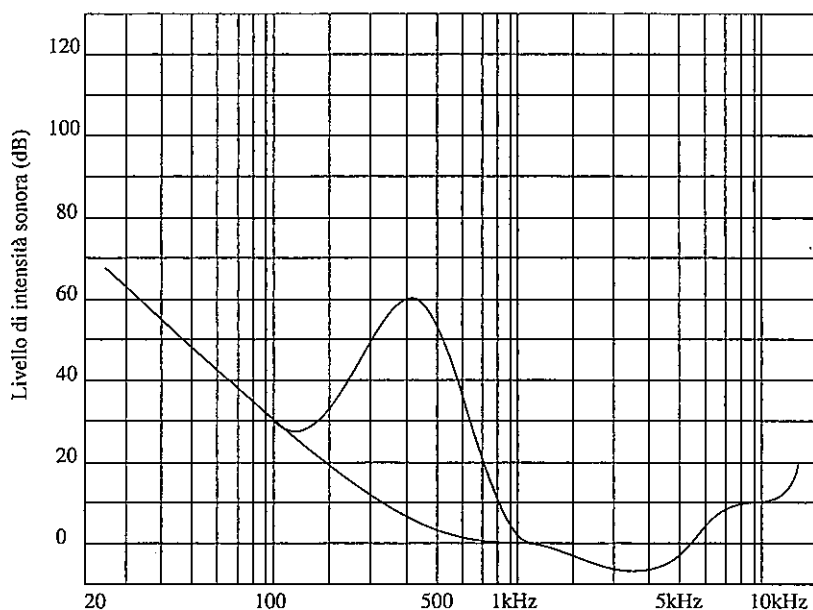



Fig. 1.8 - Soglia di udibilità dinamica.

1.3 CARATTERISTICHE DELL'APPARATO VOCALE

Nello studio del segnale vocale è opportuno considerare il parlato come concatenazione di suoni elementari, detti *fonemi*. Il numero necessario a rappresentare tutti i suoni caratteristici di un determinato linguaggio dipende dal linguaggio stesso. In particolare, per la lingua italiana sono stati catalogati 34 fonemi. Nella generazione dei differenti fonemi da parte dell'apparato vocale è necessario distinguere due aspetti che sono il tipo di sorgente utilizzata per generare il flusso d'aria che produce il fonema (eccitazione) e le trasformazioni che l'eccitazione subisce nell'attraversamento del cavo orale (fig. 1.9).

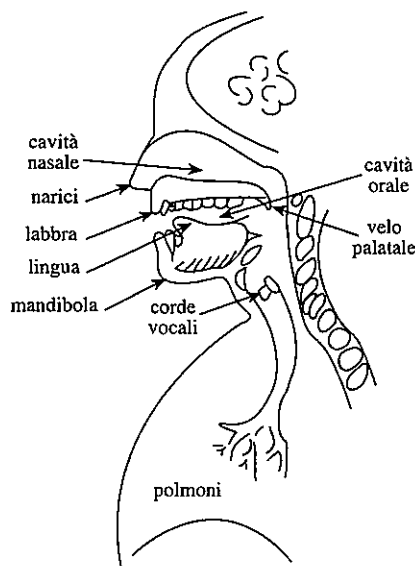



Fig. 1.9 - Tratto vocale.

A secondo del tipo di eccitazione è possibile distinguere tra fonemi vocalizzati e non vocalizzati. I fonemi vocalizzati sono legati principalmente alla pronuncia delle vocali. Come eccitazione per tali suoni si utilizzano gli impulsi prodotti dall'apertura e chiusura periodica delle corde vocali a seguito del loro attraversamento da parte dal flusso d'aria proveniente dai polmoni (fig. 1.10). Analizzando l'andamento della pressione generata del flusso d'aria prodotto dalle corde vocali, si vede che questa è una funzione periodica (indipendente

dal fonema emesso) con forma d'onda approssimativamente triangolare asimmetrica (con tempo di salita maggiore del tempo di discesa) e con un tempo di apertura che varia dal 30% al 70% del periodo T. Indicando con τ_1 ed τ_2 rispettivamente la durata del ramo crescente e decrescente della forma d'onda, una sua approssimazione analitica è ottenibile come [Rab78]

$$g(t) = \begin{cases} \frac{1}{2}[\,1 - \cos(\pi t/\tau_1)\,]\,; & 0 \le t \le \tau_1 \\ \cos[\,\pi(t-\tau_1)/2\tau_2\,]\,; & \tau_1 \le t \le \tau_1 + \tau_2 \\ 0\,; & \tau_1 + \tau_2 \le t \le T \end{cases} \qquad (1.5)$$

alla quale corrisponde uno spettro che, essendo, l'eccitazione periodica, è a righe equispaziate in frequenza di un valore pari alla fondamentale del segnale (*pitch*). A causa della differente lunghezza delle corde vocali, la fondamentale dell'eccitazione varia nell'intorno dei 125 Hz per la voce maschile, mentre ha frequenza doppia per quella femminile. Fissato il parlatore, le variazioni nella fondamentale dell'eccitazione sono principalmente legate all'intonazione. Per quanto riguarda l'inviluppo dello spettro, questo decade esponenzialmente con la frequenza di circa 12 dB/ottava.

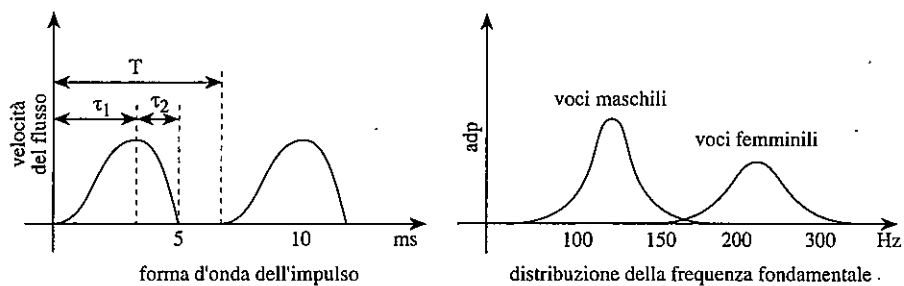



Fig. 1.10 - Caratteristiche dell'eccitazione per suoni vocalizzati.

Data la non stazionarietà del segnale vocale, è necessario precisare che l'analisi in frequenza deve essere eseguita su segmenti di segnale limitati in un intervallo temporale all'interno del quale la sorgente può essere approssimata come quasi stazionaria. La voce, in generale, è considerabile un segnale quasi stazionario se si limita l'analisi ad intervalli di una decina di ms.

Passando ai fonemi non vocalizzati, si ha che questi intervengono insieme ai fonemi vocalizzati nella pronuncia delle consonanti. Si hanno due meccanismi principali di eccitazione. Il primo sfrutta la turbolenza che si genera in corrispondenza di restringimenti della cavità orale, come nella pronuncia della consonanti fricative (es.: "s"). Il secondo è di tipo impulsivo ed utilizza il transitorio generato da interruzioni e bruschi rilasci del flusso d'aria che attraversa la cavità stessa, come nella pronuncia della consonanti occlusive (es.: "t"). In entrambi i casi il segnale generato è assimilabile a rumore bianco e presenta, quindi, uno spettro estremamente esteso.

Definiti i due differenti tipi di eccitazione comuni ai vari fonemi, questi ultimi sono generati cambiando le caratteristiche di propagazione dell'eccitazione all'interno del cavo orale ed il tipo di irradiazione utilizzata per la loro pronuncia. Il cavo orale, infatti, ha per la generazione della voce un ruolo paragonabile a quella che ha la cassa di risonanza per uno strumento musicale. A seguito dell'instaurazione di moti naturali al suo interno, si hanno esaltazioni o attenuazioni dello spettro del segnale d'eccitazione, con conseguenti differenziazioni dell'uscita prodotta. Dato che i moti naturali interessano componenti armoniche aventi lunghezze d'onda proporzionali alla distanza tra le pareti della cavità, lo spettro del segnale prodotto dipende dalla conformazione (forma e dimensioni) che la bocca assume durante la pronuncia. La conformazione della bocca è da un lato fissata dalla fisionomia del parlatore, mentre dall'altro è variabile variando la posizione della lingua, della mandibola e delle labbra. Anche l'irradiazione è regolata dal parlatore, innanzitutto bilanciando il contributo della bocca e delle narici tramite la posizione del velo palatale ed inoltre, il contributo da parte della bocca è regolato dal grado di ostruzione del cavo orale da parte della lingua e delle labbra.

Riepilogando, la differenziazione tra fonemi è ottenuta variando il tipo di eccitazione e di propagazione adottata (fig. 1.11). Passando ad analizzare nel dettaglio i differenti tipi di fonemi, si nota che nella pronuncia delle vocali si utilizzano esclusivamente i suoni vocalizzati. Come già detto, l'eccitazione utilizzata è identica per tutte le vocali ed i differenti fonemi si ottengono tramite differenti configurazioni del cavo orale. Le risonanze che si instaurano all'interno della cavità, producono effetti di esaltazione di componenti spettrali, che si manifestano come picchi nell'inviluppo dello spettro del segnale, dette *formanti*. Un suono vocalizzato, dunque, è ben caratterizzabile in frequenza. Il

suo spettro ha un inviluppo dipendente dalla posizione delle formanti, sul quale si sovrappone un'ondulazione, legata all'eccitazione, con frequenza pari alla fondamentale. È opportuno sottolineare che tale modellizzazione, per quanto sufficiente per il seguito della trattazione, risulta semplificata. Per una trattazione più accurata si rimanda alla bibliografia [Fla72].

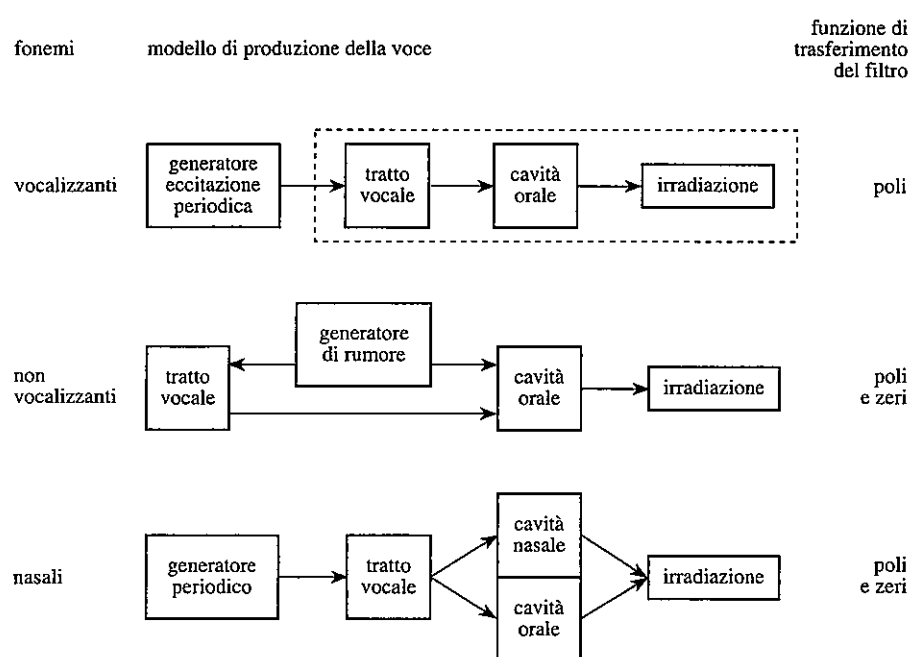



Fig. 1.11 - Modelli della sorgente del segnale vocale.

Per analizzare gli effetti sulla voce della frequenza a cui sono poste le formanti, si osserva che un arretramento della posizione della lingua o una maggiore apertura della mandibola favorisce la realizzazione di cavità più ampie all'interno della bocca: ciò permette l'instaurazione di risonanze a con lunghezza d'onda maggiore e, quindi, frequenza minore (fig. 1.12).

Riducendosi la frequenza delle formanti, il suono emesso risulta più grave. Le combinazioni di sezione della cavità orale e posizione della lingua nella pronuncia delle differenti vocali sono riportate in tabella

sezione\posizione	anteriore	centrale	posteriore
piccola	i (ira)		u (luna)
media	é (nero)		ó (volo)
grande	è (bene)	a (pane)	ò (cosa)

Tab. 1.2 - Configurazione del cavo orale per i differenti fonemi vocalizzati.

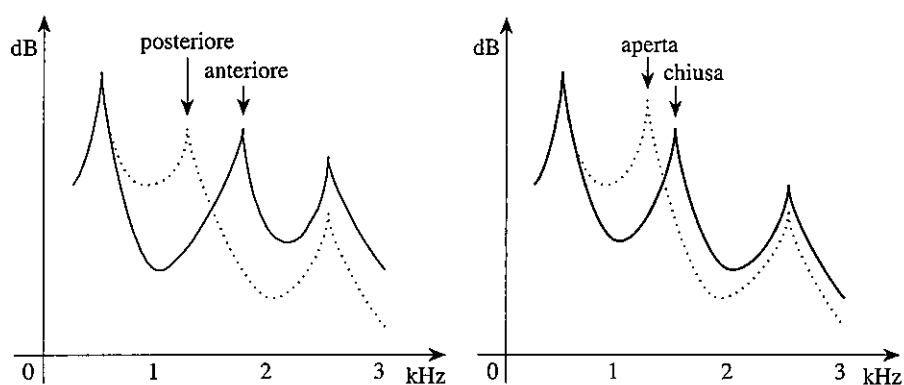



Fig. 1.12 - Influenza sulle formanti della posizione della lingua e della sezione del cavo orale.

Analizzando solamente la posizione delle prime due formanti, è possibile distinguere tra i differenti fonemi vocalizzati. Infatti, graficando il valore delle frequenze a cui sono posizionate le formanti per differenti vocali pronunciate da differenti parlatori, è possibile raggruppare i fonemi in insiemi sufficientemente disgiunti (fig. 1.13). In ogni caso, la frequenza della prima formante non scende al di sotto dei 200 Hz, mentre la frequenza massima per la seconda formante è nell'intorno dei 2500 Hz. Volendo, dunque, dimensionare il canale audio in modo tale da lasciar transitare inalterata l'informazione associata ai suoni vocalizzati, si potrebbe limitare la banda del canale a tale intervallo. Si noti come, in tal modo, si eliminerebbe la fondamentale del segnale, perdita non grave, data la capacità dell'apparato uditivo di ricavare la fondamentale a partire da sue armoniche.

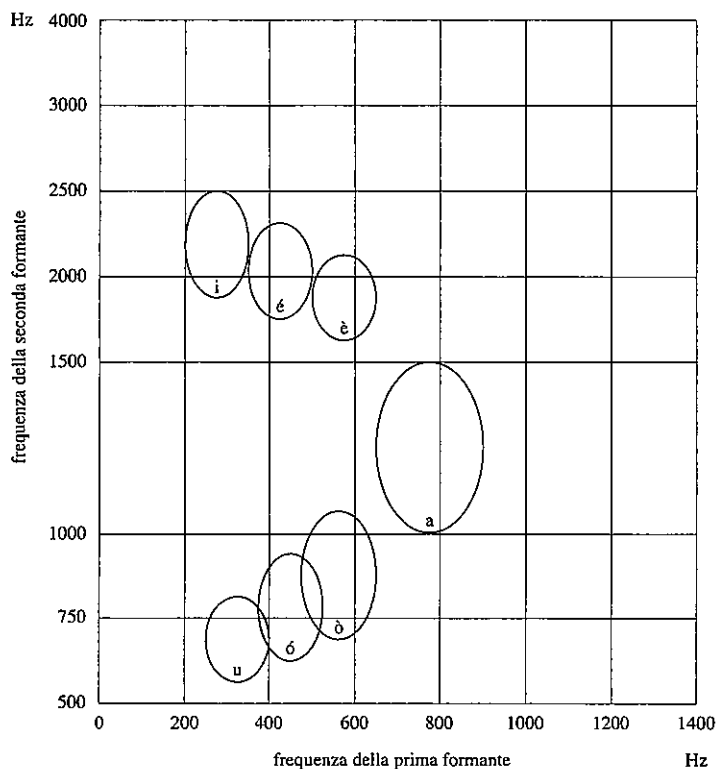



Fig. 1.13 - Posizione delle prime due formanti per suoni vocalici.

Passando all'analisi delle caratteristiche dei fonemi vocalizzati nel dominio del tempo, dato il tipo di eccitazione, essi risultano essere periodici, con periodo pari a quello dell'eccitazione stessa. Le forme d'onda risultano essere composte da treni di oscillazioni smorzate, dipendenti dalla funzione di trasferimento del cavo orale (fig. 1.14). Grazie all'inerzia della sorgente utilizzata per l'eccitazione (il flusso d'aria proveniente dai polmoni) tali segnali risultano stazionari per intervalli di tempo consistenti (circa 200 ms). Inoltre, anche grazie all'amplificazione dovuta alle risonanze in corrispondenza delle formanti, l'ampiezza dei suoni vocalizzati (e quindi la loro potenza) è elevata (es.: circa 50 mW nella pronuncia della "o").

Un modello dell'apparato vocale nel dominio del tempo per la generazione di fonemi vocalizzati può essere realizzato a partire da un filtro, avente in ingresso un generatore di segnali impulsivi periodici, con periodo pari a quello dell'eccitazione. La funzione di trasferimento del filtro è a soli poli,

posti in corrispondenza delle frequenze delle formanti. L'uscita di tale filtro va poi sottoposto ad un leggero filtraggio passa basso, con frequenza di taglio dell'ordine del kHz, per tenere conto della funzione di trasferimento dell'irradiazione da parte delle labbra [Rab78].

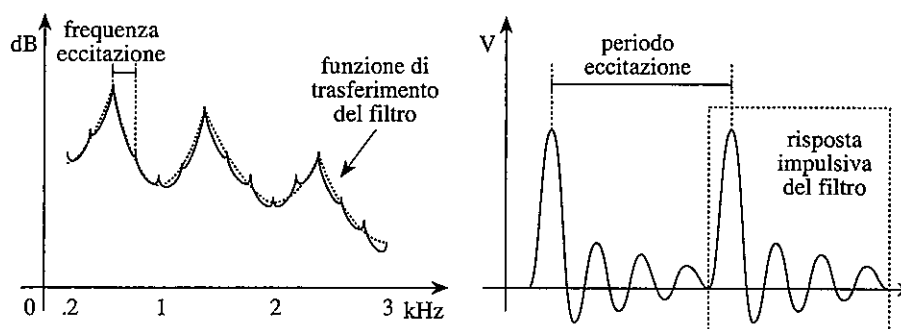



Fig. 1.14 - Contributo dell'eccitazione e della funzione di trasferimento del cavo orale sullo spettro e sulla forma d'onda del segnale vocale.

La generazione delle consonanti, invece, avviene tramite meccanismi più complessi. Innanzitutto non si ha un solo tipo di eccitazione, ma, oltre all'eccitazione tramite rumore tipico delle consonanti stesse, può essere presente anche un'eccitazione vocalizzata. In questo secondo caso, le consonanti vengono definite sonore, altrimenti sorde. Inoltre, l'irradiazione delle consonanti non avviene esclusivamente tramite le labbra, ma viene sfruttata anche l'irradiazione da parte delle cavità nasali. Una classificazione dei fonemi relativi alle consonanti deve tenere conto di come i differenti parametri (tipo di eccitazione, geometria del cavo orale, tipo di irradiazione) sono combinati.

Una prima classe di fonemi non vocalizzati sono i fricativi che, come già accennato, utilizzano come eccitazione la turbolenza che si genera in corrispondenza di restringimenti della cavità orale. Una loro classificazione è possibile in funzione della posizione di tale restringimento, distinguendo consonanti labiodentali (es.: "f" per le sorde o "v" per le sonore), dentali (es.: "s" in "sano" per le sorde o in "rosa" per le sonore), o alveolari (es.: "c" in "cena" per le sorde o "g" in "gelo" per le sonore).

Le consonanti occlusive utilizzano come eccitazione il transitorio generato da interruzioni e bruschi rilasci del flusso d'aria che attraversa la cavità

orale. Anche per le occlusive è possibile distinguere tra consonanti labiali (es.: "p" per le sorde o "b" per le sonore), alveolari (es.: "t" per le sorde o "d" per le sonore), o palatali (es.: "k" per le sorde o "g" in "gatto" per le sonore) in funzione della posizione di tale occlusione.

Un'altra classe di consonanti si ottiene utilizzando un'eccitazione di tipo vocalizzato, ma con un'irradiazione non esclusivamente ottenuta tramite le labbra. Ostruendo solo parzialmente la cavità orale con la lingua, si hanno le semivocali. A secondo della posizione della lingua stessa, si distingue tra consonanti palatali (es.: "r") ed alveolari (es.: "l"). Bloccando totalmente l'irradiazione da parte della bocca (tramite le labbra o la lingua), l'irradiazione avviene solamente tramite le cavità nasali. Tali consonanti, dette appunto nasali, si distinguono, in funzione della posizione dell'occlusione, in labiali (es.: "m") o palatali (es.: "n").

Dal punto di vista della rappresentazione in frequenza dei fonemi non vocalizzati, data l'ampiezza dello spettro della loro eccitazione, si ha che la banda richiesta (circa 10 kHz) è notevolmente più estesa di quella dei vocalizzati (fig. 1.15). Dal punto di vista dell'ampiezza, invece, dato che l'eccitazione sfrutta flussi d'aria meno consistenti di quanto avviene per suoni vocalizzati e che la funzione di trasferimento presenta degli zeri, la potenza dei relativi fonemi è tipicamente inferiore (fig. 1.16) (es.: circa 0.03 mW nella pronuncia della "v").

Un modello dell'apparato vocale per la pronuncia di consonanti è sostanzialmente differente da quello utilizzato per la pronuncia delle vocali. Innanzitutto per l'eccitazione è necessario affiancare al generatore periodico utilizzato per i suoni vocalizzati, un generatore di rumore. Inoltre, mentre la funzione di trasferimento del cavo orale per le vocali può essere approssimata da un filtro a soli poli (a causa delle risonanze), la funzione di trasferimento nel caso della pronuncia di consonanti presenta anche degli zeri a causa dei differenti tipi di irradiazione. Infatti, da un punto di vista qualitativo, l'energia necessaria per l'instaurazione di risonanze all'interno del cavo orale è da considerare persa se l'irradiazione avviene tramite le cavità nasali. Di conseguenza, alle frequenze di risonanza, la funzione di trasferimento presenta degli zeri. Dato che modelli digitali a poli e zeri portano a realizzazioni di complessità computazionale maggiore rispetto a quella di modelli a soli poli, si preferisce adottare ancora modelli a soli poli, ma di ordine maggiore di quelli utilizzati per i fonemi vocalizzati.

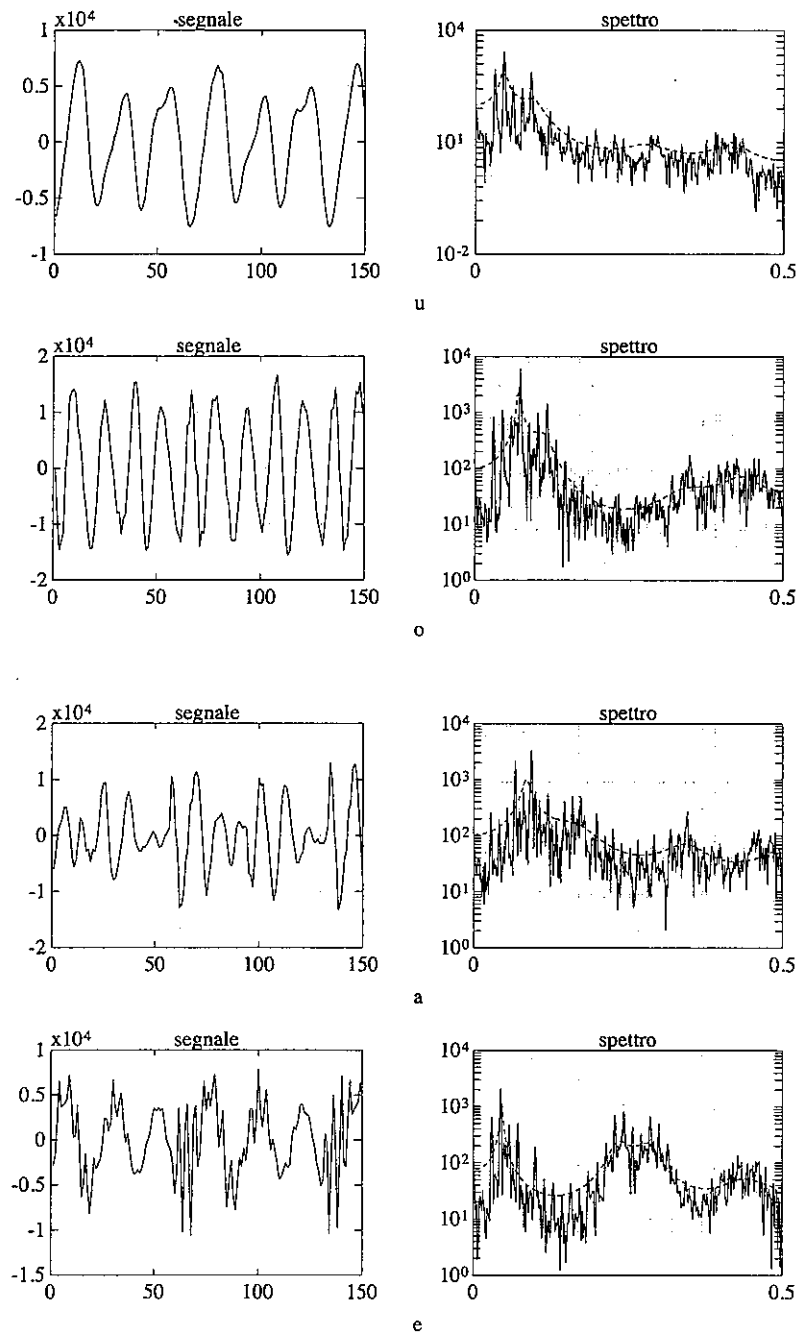



Fig. 1.15a - Forme d'onda e spettri di differenti fonemi.

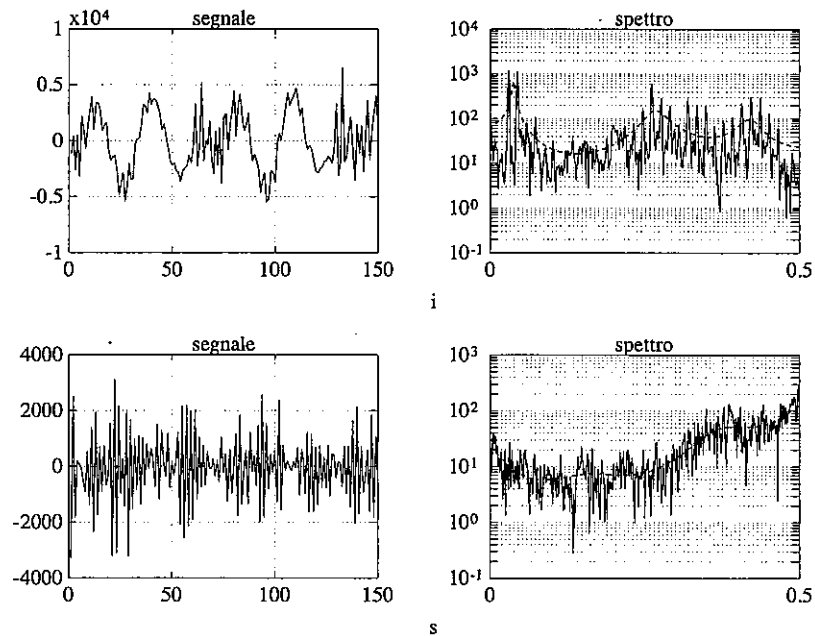



Fig. 1.15b - Forme d'onda e spettri di differenti fonemi.

Riepilogando, i suoni vocalizzati sono caratterizzati da forme d'onda periodiche con banda limitata a pochi kHz, di notevole ampiezza e durata. I suoni non vocalizzati sono caratterizzati da forme d'onda irregolari, con una banda superiore ai 10 kHz, ma di ampiezza e durata tipicamente inferiore a quelle dei vocalizzati.

1.4 TRASDUZIONE ELETTROACUSTICA

La trasmissione di segnali audio richiede, innanzitutto, una trasformazione del segnale da variazioni della pressione dell'aria in funzione del tempo in un segnale elettrico analogico. A destinazione, il segnale elettrico è poi riconvertito in un segnale audio. Tali trasformazioni avvengono tramite *trasduttori elettroacustici*. I trasduttori che eseguono la trasformazione acustico-elettrica vengono detti *microfoni*, *altoparlanti* quelli utilizzati nella trasformazione elettrico-acustica. Un trasduttore è detto reversibile se è in grado

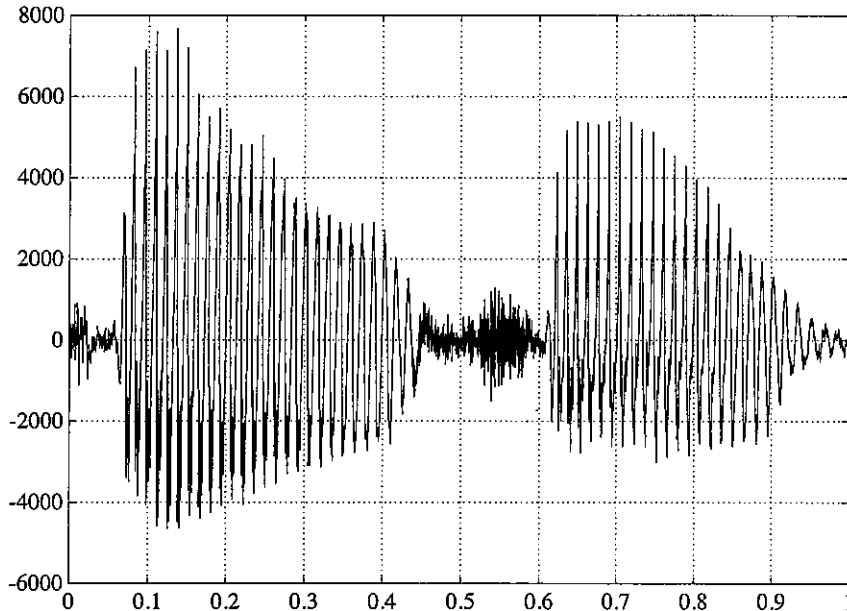



Fig. 1.16 - Forma d'onda della parola "casa": si confrontino le ampiezze dei fonemi vocalizzati e non vocalizzati.

di lavorare sia come altoparlante che come microfono, non reversibile altrimenti. Un esempio di trasduttore non reversibile è il microfono a carbone. In esso, il segnale acustico che incide sul trasduttore comprime i granuli di carbone contenuti in una capsula. Ciò fa variare la resistenza che è presentata ai capi di due elettrodi, modulando, così la corrente che lo attraversa. Attualmente, i trasduttori utilizzati sono quasi esclusivamente trasduttori reversibili, per le loro migliori caratteristiche.

La conversione acustico-elettrica è, tipicamente, ottenuta dalla sequenza di una conversione acustico-meccanica e di una conversione meccanico-elettrica. La prima converte variazioni di pressione in movimento indotto in struttura meccanica mobile, normalmente utilizzando una membrana. La funzione di trasferimento di questa prima trasformazione è legata alla geometria e alle proprietà dei materiali utilizzati per il supporto e per la membrana.

A secondo dei principi secondo i quali viene eseguita la conversione meccanico-elettrica, i trasduttori vengono distinti in magnetici, elettrostatici e piezoelettrici. Tra i magnetici si trovano i più diffusi trasduttori elettroacustici,

che sono gli elettrodinamici (fig. 1.17). Nel seguito viene data, a titolo di esempio, una breve descrizione del funzionamento di tali trasduttori. Essi generano un segnale elettrico sfruttando le variazioni di flusso elettromagnetico che si hanno all'interno di una bobina, a seguito del movimento di quest'ultima. Negli altoparlanti, il movimento della membrana è indotto dal segnale che attraversa la bobina. Indicando con $\boldsymbol{i}$ il valore della corrente iniettata nella bobina, con l la lunghezza dell'avvolgimento e con $\boldsymbol{B}$ il vettore induzione magnetica, la forza $\boldsymbol{F}$ che si genera nell'avvolgimento è pari a

$$\mathbf{F} = \mathbf{B} \times \mathbf{i} \cdot \mathrm{l} \tag{1.6}$$

Nei microfoni tale movimento è generato dalle variazioni di pressione che si hanno sulla membrana, alla quale la bobina è solidale. In tal caso la f.e.m. $\boldsymbol{E}$ che si genera ai morsetti dell'avvolgimento è pari a

$$\mathbf{E} = \mathbf{B} \cdot \mathbf{v} \cdot \mathrm{l} \tag{1.7}$$

dove $\boldsymbol{v}$ rappresenta il vettore velocità della bobina. Essa è legato alla forza $\boldsymbol{F}$ applicata alla bobina tramite la relazione

$$\mathbf{v} = \frac{\mathbf{F}}{\underline{Z}_m} \tag{1.8}$$

La costante complessa $\underline{Z}_m$ è detta impedenza meccanica ed è legata alla resistenza che la parte meccanica oppone al movimento. Scomponendola nella sua parte reale ed immaginaria

$$\underline{Z}_m = R_m + j\,X_m(\omega) \tag{1.9}$$

si ha che la componente reale Rm è legata essenzialmente agli effetti dissipativi dovuti alla flessione del materiale che costituisce l'ancoraggio dell'equipaggiamento mobile. In realtà essa non risulta indipendente dalla frequenza, ma cresce leggermente all'aumentare della stessa. Per quanto riguarda la X_m, legata all'inerzia della meccanica, può essere ulteriormente scomposta come

$$X_m = \omega m - \frac{s}{\omega} \tag{1.10}$$

dove la prima componente è dovuta alla massa m dell'equipaggio mobile ed aumenta all'aumentare della frequenza del segnale, al contrario della seconda componente, dovuta alla rigidità s del materiale. Il valore dell'impedenza meccanica, e quindi la funzione di trasferimento del trasduttore acustico- meccanico, non è, dunque, lineare in frequenza, ma presenta di un minimo (fig. 1.17). Ciò crea problemi, oltre che per la non linearità della risposta in frequenza, anche per l'esaltazione dei comportamenti non lineari dei materiali utilizzati a seguito dell'incremento dell'ampiezza delle escursioni della bobina per segnali a frequenza prossima a quella su cui si trova tale minimo.

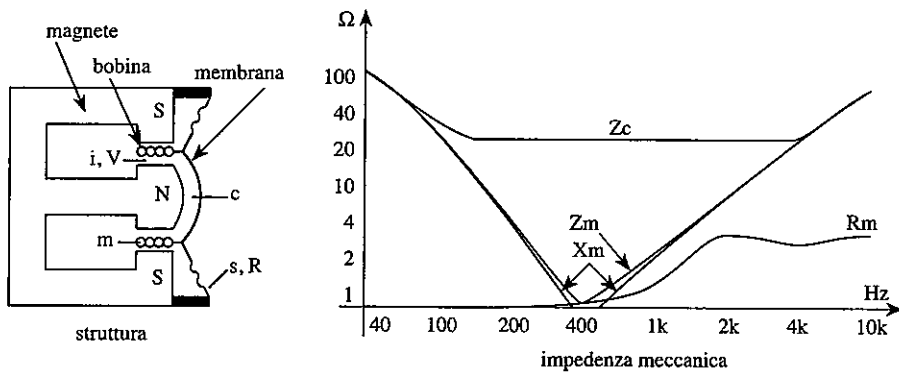



Fig. 1.17 - Trasduttori elettrodinamici.

Per migliorare la linearità in frequenza del trasduttore, si potrebbe pensare di rendere predominante la componente dissipativa Rm. Un incremento dell'impedenza meccanica a parità di forza incidente, però, comporterebbe una riduzione della tensione generata nel caso di microfoni e, analogamente, ridurrebbe l'efficienza nel caso di altoparlanti. La soluzione generalmente adottata è, invece, quella di compensare meccanicamente il trasduttore, sfruttando la dinamica delle masse d'aria presenti al suo interno. In tal modo si ottiene un nuovo valore $\mathbf{Z_c}$ dell'impedenza meccanica

$$\underline{Z}_c = (R+V) + j\left[(m+i)\,\omega - \frac{s+c}{\omega}\right] \tag{1.11}$$

dove la resistenza R è corretta tramite le perdite per attrito viscoso V che si hanno nella camera posteriore del supporto, la componente dell'impedenza

meccanica dovuta alla massa "m" è corretta sfruttando l'inerzia "i" del flusso d'aria generato dal movimento della membrana, mentre la rigidità "s" è corretta tramite un coefficiente "c" che tiene conto del comportamento elastico alla compressione dell'aria da parte della membrana (fig. 1.17). Ulteriori interventi sulla risposta in frequenza si possono ottenere sfruttando opportunamente le risonanze generate nella camera posteriore del supporto.

Per quanto riguarda le prestazioni dei trasduttori elettrodinamici, questi hanno caratteristiche molto buone in termini di linearità della risposta in frequenza, ma sono caratterizzati da impedenze modeste (tipicamente 8 W), il che lirende utilizzabili solamente in sistemi audio elettronici amplificati. Sempre della famiglia dei trasduttori elettromagnetici sono quelli a nastro (fig. 1.18). In essi l'elemento mobile è rappresentato da un sottile nastro conduttore corrugato posto tra due espansioni polari. Mentre le caratteristiche dei trasduttori a nastro in termini di risposta in frequenza sono tra le migliori ottenibili, le caratteristiche elettriche sono ulteriormente peggiorate rispetto agli elettrodinamici (impedenza di qualche frazione di Ω). Ciò li rende idonei essenzialmente per applicazioni professionali.

Passando ai trasduttori elettrostatici (fig. 1.18), i segnali in gioco sono legati alla variazione di campo elettrico dovute al movimento di una membrana metallica rispetto ad un elettrodo fisso e quindi ad una modulazione della capacità del trasduttore. A secondo che la polarizzazione degli elettrodi sia mantenuta tramite un generatore esterno o sia dovuta alla polarizzazione permanente del dielettrico che separa le due armature, i trasduttori elettrostatici si distinguono, rispettivamente, in trasduttori a condensatore o a elettrete. Indicando con "d" la distanza tra gli elettrodi, con "S" la loro superficie e con "ε_0" la costante dielettrica, la capacità formata dalle due armature e la carica in essa immagazzinata sono pari a

$$C = \frac{\varepsilon_0 S}{d} ; \quad Q = C V_p = \frac{\varepsilon_0 S V_p}{d} \tag{1.12}$$

dove Vp rappresenta la tensione di polarizzazione del dielettrico. Ipotizzando che la carica sulle armature si mantenga costante, la tensione V generata ai capi di quest'ultime a seguito di un loro spostamento $\xi(t)$ è pari a

$$V(t) = \frac{Q}{\varepsilon_0 S} [d + \xi(t)] \tag{1.13}$$

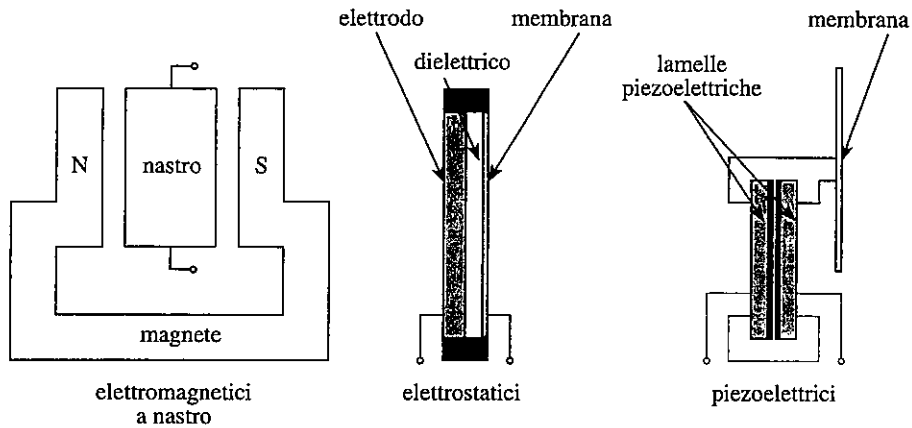



Fig. 1.18 - Esempi di trasduttori elettroacustici reversibili.

La qualità del segnale ottenuto da questi trasduttori è elevata, paragonabile a quella dei trasduttori a nastro. Data, però, l'impedenza elevata che li caratterizza (dell'ordine del MΩ), l'ampiezza dei segnali generati da microfoni a condensatore è modesta, il che rende indispensabile uno stadio di preamplificazione, a volte integrato nel trasduttore stesso.

I trasduttori piezoelettrici (fig. 1.18), invece, sfruttano la caratteristica di alcuni cristalli (essenzialmente quarzi, ceramiche piezoelettriche o polifluoruro di vinile) di deformarsi se sottoposte ad un campo elettrico e, viceversa, di generare un campo elettrico se sottoposti a deformazioni meccaniche. Tale trasduzione diretta meccanico-elettrica è caratterizzata, però, da ampiezze delle deformazioni estremamente modeste. Affinché il livello delle deformazioni raggiunga livelli sfruttabili in applicazioni commerciali, si ricorre all'accoppiamento di strati di materiale con caratteristiche complementari, in grado di eseguire un'amplifi- cazione meccanica delle deformazioni. Ciò può essere ottenuto, ad esempio, accoppiando due lamine piezoelettriche polarizzate in senso opposto, in modo tale che, a fronte di un campo elettrico, mentre una tende a contrarsi, l'altra si espande. A causa delle buone caratteristiche elettriche e della notevole robustezza, tali trasduttori sono diffusissimi per applicazioni commerciali.

In tabella sono riepilogate le caratteristiche tipiche dei principali trasduttori in termini di sensibilità ed impedenza, dove la prima grandezza è misurata

come il rapporto in dB tra la tensione E generata (espressa in volt) a fronte di un segnale con una pressione P (espressa in microbar) e frequenza di 1 kHz.

Tipo	Sensibilità (dB)	Impedenza (Ω)
A carbone	-45	100
Elettrodinamico	-85	10
A nastro	-105	1
A condensatore	-50	1000000
Piezoelettrico	-50	100000

Tab. 1.3 - Caratteristiche dei trasduttori elettroacustici.

1.5 IL CANALE AUDIO E TELEFONICO

Affinché il canale audio sia in grado di trasmettere fedelmente un qualsiasi segnale, le sue caratteristiche in termini di banda e rapporto segnale rumore debbono essere migliori della banda e della dinamica apprezzabili dall'udito. In tal caso il segnale ricevuto a destinazione risulterebbe indistinguibile dall'originale emesso dalla sorgente (audio HiFi). Dall'analisi precedentemente fatta sulle caratteristiche dell'apparato uditivo si ricava che la banda richiesta in tal caso al canale audio è di circa 20 kHz e la dinamica di 120 dB. Mentre i requisiti di banda non rappresentano attualmente un limite implementativo, ma solo economico, la dinamica richiesta risulta elevata. Infatti, i sistemi HiFi analogici commerciali forniscono circa 60 dB di dinamica, mentre i sistemi digitali (Compact Disk e Digital Audio Tape) sfiorano i 100 dB. Nei sistemi di radiodiffusione sia analogici (FM) che numerici (Digital Satellite Radio), dove la banda ha un peso maggiore, questa viene ridotta a 15 kHz.

Per quanto riguarda il segnale telefonico, le specifiche sul canale (banda e dinamica) possono essere meno stringenti, viste le caratteristiche della sorgente. Per quanto riguarda la banda, la maggiore ampiezza dei suoni vocalizzati fa si che la potenza media del segnale vocale sia concentrata essenzialmente nella parte inferiore dello spettro [ITU-T P.50] (fig. 1.20). Analizzando in dettaglio l'andamento dello spettro a lungo termine, infatti, è evidente il contributo delle prime armoniche dell'eccitazione. La banda utile del segnale è, quindi, limitata a meno di 4 kHz, permettendo in tal modo una

sia pur parziale riproduzione dei suoni non vocalizzati, indispensabili per una buona intelligibilità della voce e per il riconoscimento del parlatore. Per la precisione, scelto un riferimento ad 800 Hz, l'attenuazione in frequenza del canale telefonico all'interno di una banda di 4 kHz deve soddisfare la maschera riportata in figura 1.19 [ITU-T G.132]. Si nota come la banda utile è in realtà limitata all'interno dei 300-3400 Hz, per garantire opportuni intervalli di transizione ai filtri di canale.

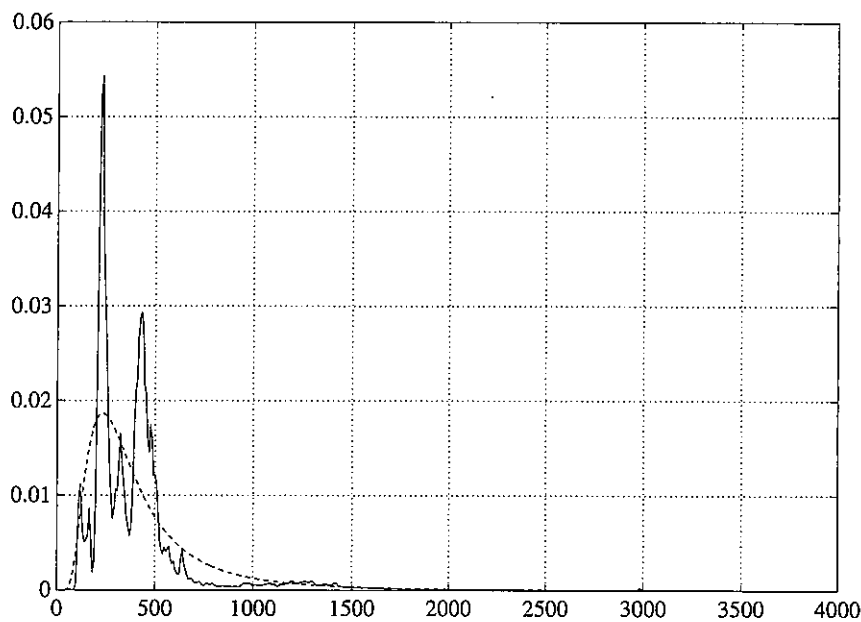



Fig. 1.19 - Densità spettrale di potenza del segnale e ITU-T P.50.

Per fissare la dinamica del canale telefonico è necessario analizzare la distribuzione statistica dell'andamento dell'ampiezza del segnale in funzione del tempo (fig. 1.21). I risultati sono fortemente influenzati dalla durata delle osservazioni. Nel caso di analisi a lungo termine (brani della durata dell'ordine del minuto) si osserva che l'ampiezza del segnale telefonico ha una distribuzione approssimativamente esponenziale [ITU-T P.50]. Ciò è giustificabile della presenza di lunghe sequenze di campioni di ampiezza modesta in corrispondenza di pause nel discorso, che aumentano la frequenza di campioni di ampiezza prossima allo zero. È possibile definire il livello massimo (o di

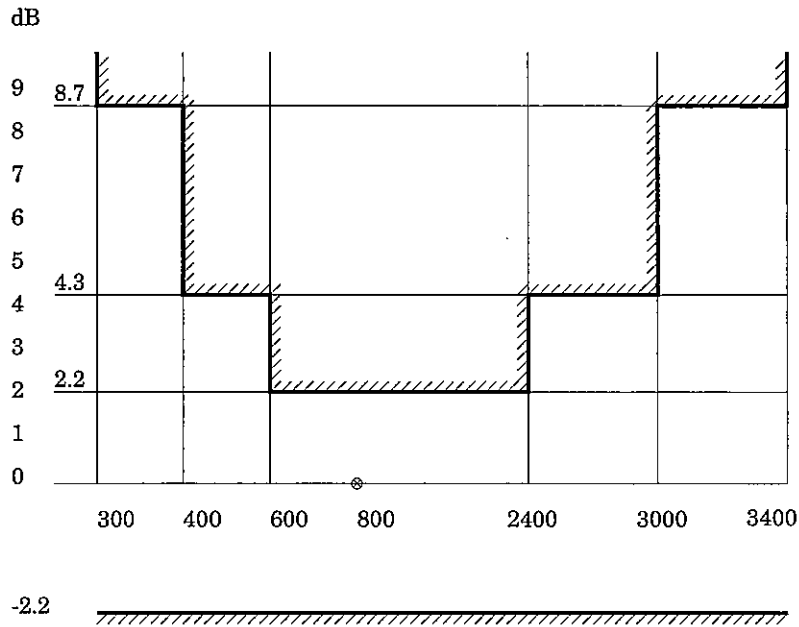



Fig. 1.20 - Maschera del canale telefonico dalla raccomandazione CCITT G.132.

picco) del segnale come il valore che viene superato in meno del 0.01% dei casi. Tale valore è estremamente variabile e si riduce progres- sivamente negli anni con il migliorare della qualità delle reti [Bon91]. Nei sistemi numerici, il livello di picco è standardizzato in 3.14 dBm0 [ITU-T G.711]: data una linea a 600 Ω, l'ampiezza corrispondente del segnale è ottenibile come

$$10 \log \frac{P_{max}}{1 \text{ mW}} = 3.14 \text{ dBm0} \rightarrow P_{max} = \frac{x_{eff}^2}{R} = 2.06 \text{ mW}$$

$$\rightarrow x_{eff} = \frac{x_{max}}{\sqrt{2}} = 1.112 \text{ V} \rightarrow x_{max} = 1.57 \text{ V} \quad (1.14)$$

A fronte di tale valore di picco, scegliendo come valore medio della potenza del segnale un livello di - 23.4 dBm0 [Bon91], si ha un valore efficace di ampiezza pari a

$$10 \log \frac{P}{1 \text{ mW}} = -23 \text{ dBm0} \rightarrow P = \frac{x_{eff}^2}{R} = 0.005 \text{ mW} \rightarrow x_{eff} = 0.055 \text{ V} \quad (1.15)$$

Per quanto riguarda la dinamica utile, essa assume un valore di circa 50 dB. In realtà in telefonia, come descritto nel seguito, la dinamica fissata per i sistemi numerici è notevolmente maggiore (> 70 dB). Ciò è necessario per garantire un certo margine per bilanciare la degradazione del segnale alla quale si va incontro nelle conversioni da numerico ad analogico (e viceversa) richieste dall'attraversamento di aree della rete a differente tecnologia.

Nel caso di analisi a breve termine si osserva che per brani della durata dell'ordine del secondo (brevi frasi), l'ampiezza del segnale ha una distribuzione approssimativamente gaussiana. Tale distribuzione rimane valida anche per brani della durata dell'ordine del millisecondo (fonemi), indipendentemente dalla natura vocalizzata o meno del segnale.

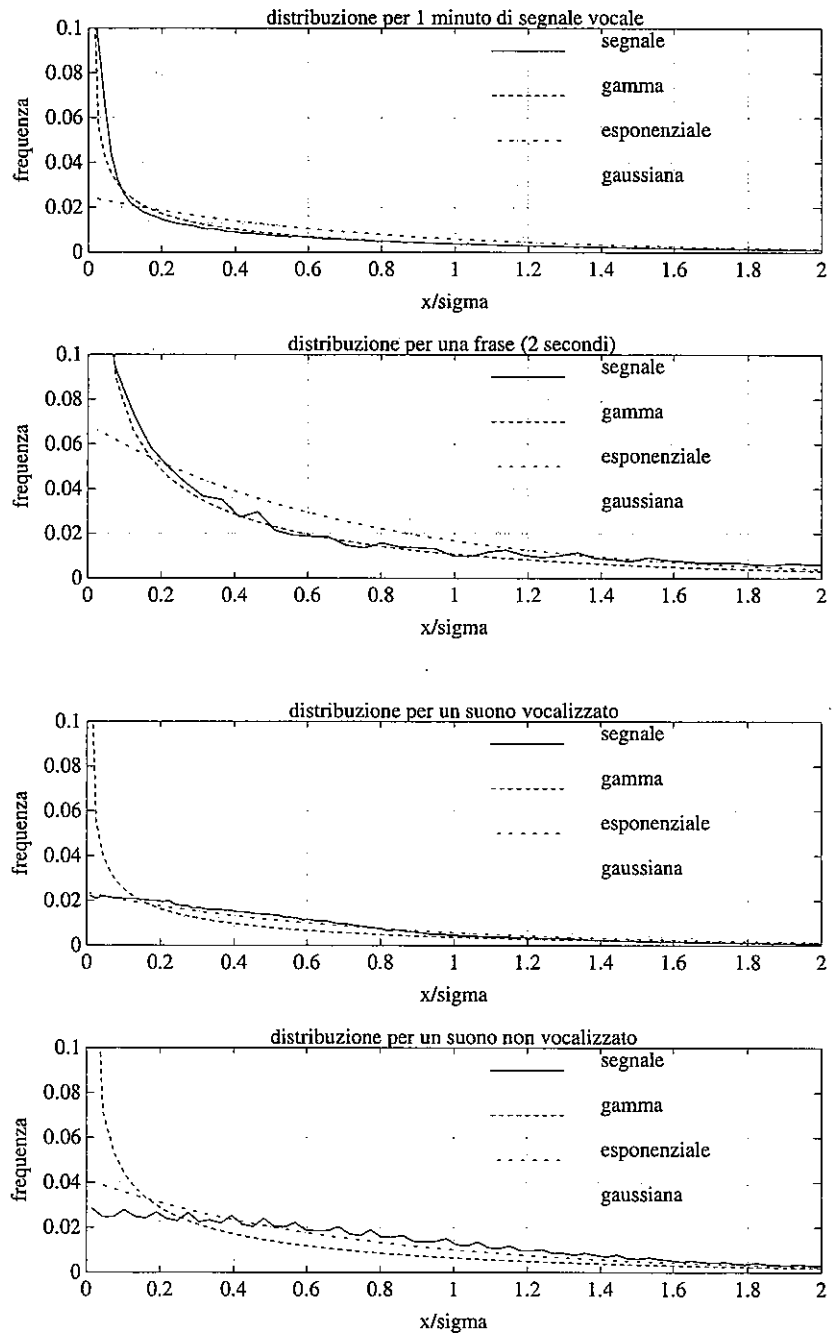



Fig. 1.21 - Distribuzione delle ampiezze per differenti tipi di segnali.

2

RAPPRESENTAZIONE NUMERICA DEI SEGNALI

2.1 GENERALITÀ SULLA CODIFICA DI SORGENTE

Un segnale analogico per poter essere trasmesso numericamente deve essere sottoposto innanzitutto ad una conversione analogico/digitale (A/D). La rappresentazione numerica ottenuta, può poi essere sottoposta a codifica, al fine di renderla più idonea alla trasmissione. A destinazione, il flusso numerico corrispondente al segnale viene riconvertito in forma analogica tramite un'eventuale decodifica ed una conversione digitale/analogica (D/A). Per quanto riguarda la codifica, si distingue tra codifica di sorgente (che ha lo scopo di comprimere il flusso numerico ottenuto dalla conversione), codifica di canale (che ha lo scopo di rendere tale flusso idoneo ad essere trasmesso su di un canale tipicamente affetto da errori) e codifica di linea (che trasforma il flusso numerico in opportuni segnali elettrici). In questo ambito si parlerà esclusivamente di codifica di sorgente. Affinché le operazioni di conversione e codifica mantengano l'informazione del segnale, è necessario comprenderne l'effetto, al fine di definire una loro implementazione ottimale. Per fornire una visione d'insieme, nel seguito si anticipano brevemente la funzione e gli effetti delle singole trasformazioni, rimandando per un'analisi più dettagliata ai paragrafi successivi.

La conversione A/D avviene concettualmente tramite due fasi di campionamento e quantizzazione (fig. 2.1). Il campionamento trasforma un segnale continuo in una serie di impulsi (campioni) equispaziati nel tempo di

ampiezza pari a quella del segnale. La rappresentazione in frequenza di un segnale campionato è dato dalla ripetizione dello spettro del segnale continuo nell'intorno di multipli della frequenza di campionamento. In ricezione, è possibile riottenere il segnale continuo dai suoi campioni eliminando tali repliche, tramite un opportuno filtraggio passa basso (interpolazione). Affinché l'interpolazione sia possibile, è necessario che le repliche non si sovrappongano allo spettro del segnale continuo e ciò è garantito se la frequenza di campionamento è pari almeno al doppio della banda del segnale. La frequenza di campionamento, dunque, viene definita in funzione della banda del segnale. Per segnali non a banda limitata, è necessario che anche l'ingresso, come l'uscita, sia sottoposto ad un filtraggio passa basso (anti-aliasing), con una frequenza di taglio pari alla metà della frequenza di campionamento.

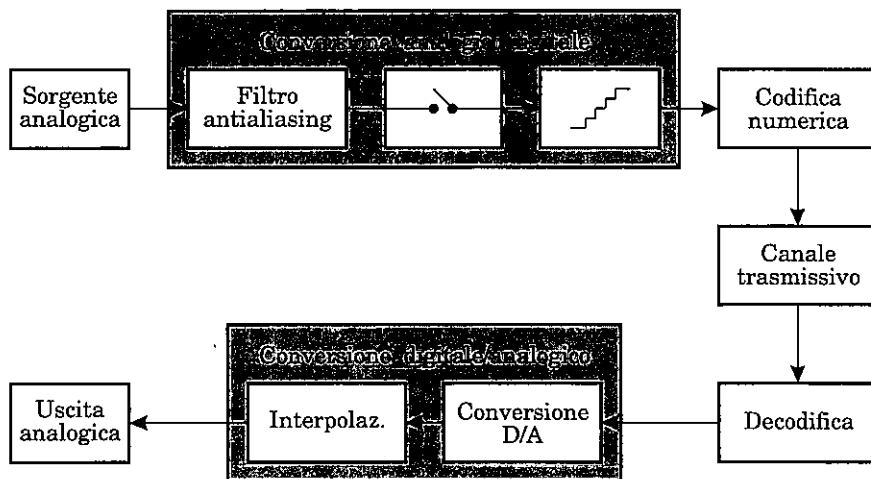



Fig. 2.1 - Struttura del canale audio numerico.

La quantizzazione permette la rappresentazione numerica dell'ampiezza dei campioni, sostituendo l'informazione sull'ampiezza del segnale con quella relativa all'appartenenza dell'ampiezza stessa a sottointervalli, detti *quanti*, definiti entro determinati estremi, detti *estremi di saturazione*. Associando a ciascun quanto un codice numerico, si ottiene la conversione A/D del segnale. In ricostruzione, in corrispondenza di ciascun quanto viene generato un segnale con ampiezza che, tipicamente, è paria al valor medio del quanto stesso. Con

la quantizzazione il segnale in ingresso viene irrimediabilmente distorto, dato che si associa ad un intervallo di valori della grandezza in ingresso un unico valore di uscita, imponendo, così, che i campioni assumano ampiezze discrete. Ciò può essere visto come una compressione tramite riduzione di entropia della sorgente. Dal punto di vista degli effetti sul segnale, la quantizzazione è interpretabile come la somma all'ingresso di un segnale di rumore, detto rumore di quantizzazione, dato dalla differenza tra l'ingresso e l'uscita. Il livello del rumore dipende da quanto grossolana è l'approssimazione ottenuta con la quantizzazione che, a sua volta, dipende dal numero di livelli di quantizzazione e quindi dal numero di bit richiesto per una etichettatura univoca dei codici. Il numero di bit utilizzati nella codifica, dunque, è legato alla degradazione imposta al segnale e quindi al rapporto segnale rumore voluto.

Fissata in funzione della banda del segnale la frequenza di campionamento e quindi il numero di campioni al secondo f_s, e fissato il numero di bit per campione R in funzione del rapporto segnale/rumore desiderato, rimane fissata la velocità f_b del flusso numerico prodotto dai sistemi di conversione A/D e D/A

$$f_b = f_s \times R \quad \text{(bit/s)} \tag{2.1}$$

Analizzando la banda necessaria per trasmettere numericamente un segnale si nota che tale codifica risulta essere estremamente inefficiente. La banda, infatti, risulta di un ordine di grandezza superiore a quella richiesta per la trasmissione analogica del segnale. D'altra parte, l'informazione presente in tale flusso numerico risulta essere estremamente ridondante e quindi comprimibile. Per qualsiasi segnale che non sia rumore bianco, infatti, i campioni non risultano indipendenti per il fatto di essere frutto dalla stessa sorgente e quindi frutto di una qualche legge di generazione (*ridondanza*). Inoltre, non considerando le caratteristiche della destinazione, non si fa nessuna analisi sulla "utilità" dei campioni stessi (*irrilevanza*). In particolare, le principali cause di ridondanza (fig. 2.2) possono essere ricondotte a:

- disuniformità della distribuzione delle ampiezze: dato che, tipicamente, la distribuzione delle ampiezze del segnale non è uniforme, è possibile ridurre la precisione del quantizzatore (e quindi il numero dei livelli) per gli intervalli di minor interesse;

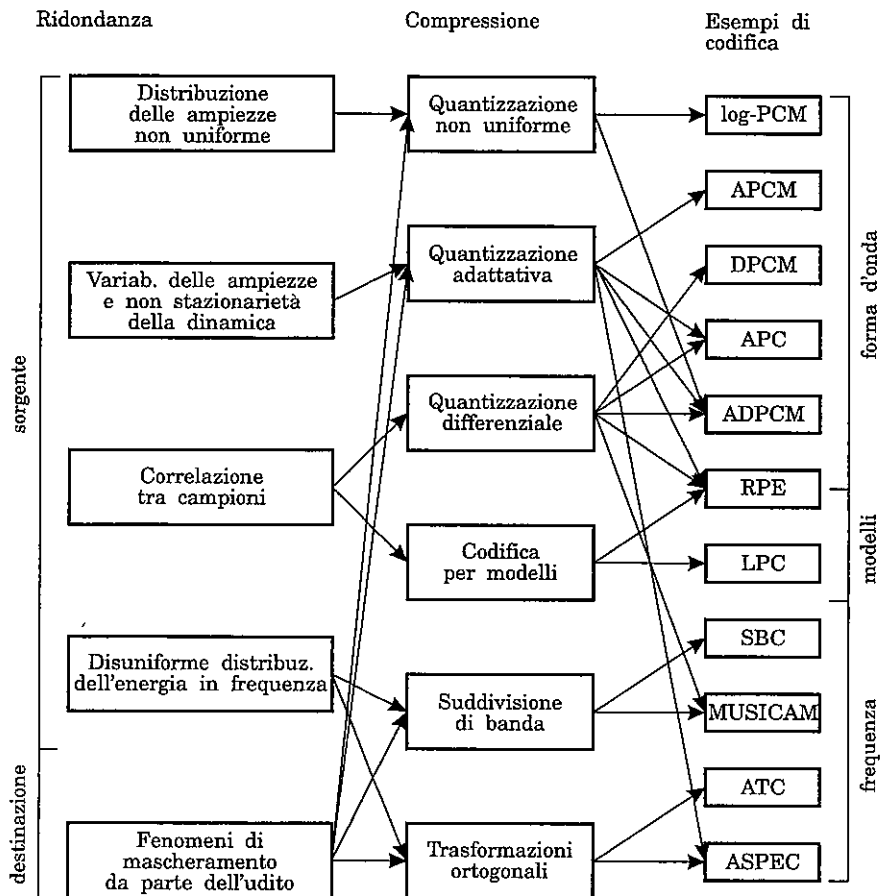



Fig. 2.2 - Tecniche di compressione e relative codifiche.

- variabilità della distribuzione delle ampiezze: la dinamica del segnale varia considerevolmente nel tempo, per cui è conveniente adottare estremi di quantizzazione che non siano costanti, ma si adattino alla dinamica corrente;
- correlazione tra campioni: l'ampiezza di un campione non è indipendente dalla serie dei campioni che lo hanno preceduto, in quanto tutti generati dalla stessa sorgente, per cui non è conveniente codificarli isolatamente;

- disuniformità della distribuzione dell'energia in frequenza: lo spettro del segnale non è uniforme, per cui è possibile adottare caratteristiche di quantizzazione differenti per ciascuna porzione dello spettro;
- fenomeni di mascheramento dell'udito: dato che la sensibilità dell'orecchio è influenzata dalle caratteristiche del suono, è possibile adottare tecniche di quantizzazione che permettano di mascherare tramite il segnale il rumore di quantizzazione.

Le tecniche di codifica in grado di eliminare tali ridondanze si possono classificare in tre gruppi, a seconda delle informazioni utilizzate per comprimere il segnale; si distingue tra:

- *codifica di forma d'onda*: tale codifica sfrutta solamente l'informazione dell'andamento in funzione del tempo del segnale da codificare. Non facendo (solitamente) riferimento al tipo di sorgente o destinazione, essa risulta applicabile ad un qualsiasi tipo di segnale;
- *codifica per modelli*: la codifica non è relativa ai campioni del segnale, ma ai parametri di un modello della sorgente in grado di generarli;
- *codifica nel dominio della frequenza*: la codifica avviene dopo una trasformazione del segnale in un dominio differente da quello del tempo, tipicamente in frequenza.

Per quanto riguarda la codifica di forma d'onda, non potendo ridurre la frequenza di campionamento, è necessario comprimere il segnale riducendo il numero dei bit per campione e quindi riducendo opportunamente il numero dei livelli del quantizzatore (fig. 2.3). Una prima classificazione delle tecniche di codifica di forma d'onda può essere fatta a seconda che esse lavorino su singoli campioni (codifica di forma d'onda senza memoria) o su blocchi di essi (codifica di forma d'onda con memoria). La codifica di forma d'onda senza memoria può avvenire solamente tramite *quantizzazione non uniforme*. In tal caso, fissati gli estremi di quantizzazione, il numero dei livelli del quantizzatore (e quindi il numero di bit della codifica) viene ridotto adottando "dove possibile" passi di quantizzazione maggiori rispetto alla quantizzazione uniforme. Il passo di quantizzazione rimane, comunque, fisso nel tempo. Per quanto riguarda la codifica di forma d'onda con memoria, essa opera sugli estremi di saturazione del quantizzatore. Si distingue tra:

- *quantizzazione adattativa*: pur adottando una quantizzazione uniforme, si riducono gli estremi di quantizzazione (e quindi il numero dei livelli del quantizzatore) rendendoli contemporaneamente variabili nel tempo in funzione della dinamica corrente del segnale. In tal modo per segnali di ampiezza modesta si utilizzano passi di quantizzazione della stessa ampiezza di quelli adottati con una quantizzazione uniforme, ma, al crescere dell'ampiezza del segnale, crescono in ampiezza mantenendo costante sia il numero di livelli che il rapporto segnale rumore;
- *quantizzazione differenziale*: invece del segnale viene codificata una grandezza da esso derivata (tipicamente la differenza tra il segnale ed una sua stima) caratterizzata da una dinamica inferiore rispetto a quella dal segnale stesso. A parità di passo di quantizzazione è così possibile ridurre gli estremi di quantizzazione, riducendo il numero dei livelli del quantizzatore.

La stima del segnale (predizione) necessaria per la quantizzazione differenziale apre la strada alla codifica per modelli. Infatti, la predizione comporta automaticamente l'identificazione di un modello della sorgente in modo tale che, disponendo di un blocco di campioni, sia possibile stimarne la legge di generazione. Ricavata, però, tale legge si dispone di un modello della sorgente. Potrebbe, quindi, risultare conveniente trasmettere i parametri che identificano il modello, piuttosto che codificare la sua uscita. Per capire quanto vantaggiosa potrebbe essere tale tecnica, si pensi, ad esempio, al problema della codifica di segnali sinusoidali. In tal caso, una singola trasmissione dei parametri che identificano il generatore (ampiezza, frequenza e fase iniziale) eliminerebbe la necessità di trasmettere l'infinita serie dei campioni della sua uscita.

Per quanto riguarda la codifica nel dominio della frequenza, essa è motivata dal fatto che analizzare il segnale in un dominio differente da quello del tempo permette di eliminare della ridondanza/irrilevanza difficilmente evidenziabile altrimenti. In particolare i fenomeni che si vogliono sfruttare sono la disuniformità dello spettro del segnale e gli effetti di mascheramento da parte dell'udito. In entrambi i casi, la compressione avviene analizzando il segnale in distinti intervalli di frequenza ed eseguendo per ciascuna banda la quantizzazione più opportuna. In particolare si tenta di ridurre il numero dei livelli del quantizzatore in funzione dell'energia contenuta nelle singole bande, sia tramite una quantizzazione adattativa, sia aumentando l'ampiezza dei

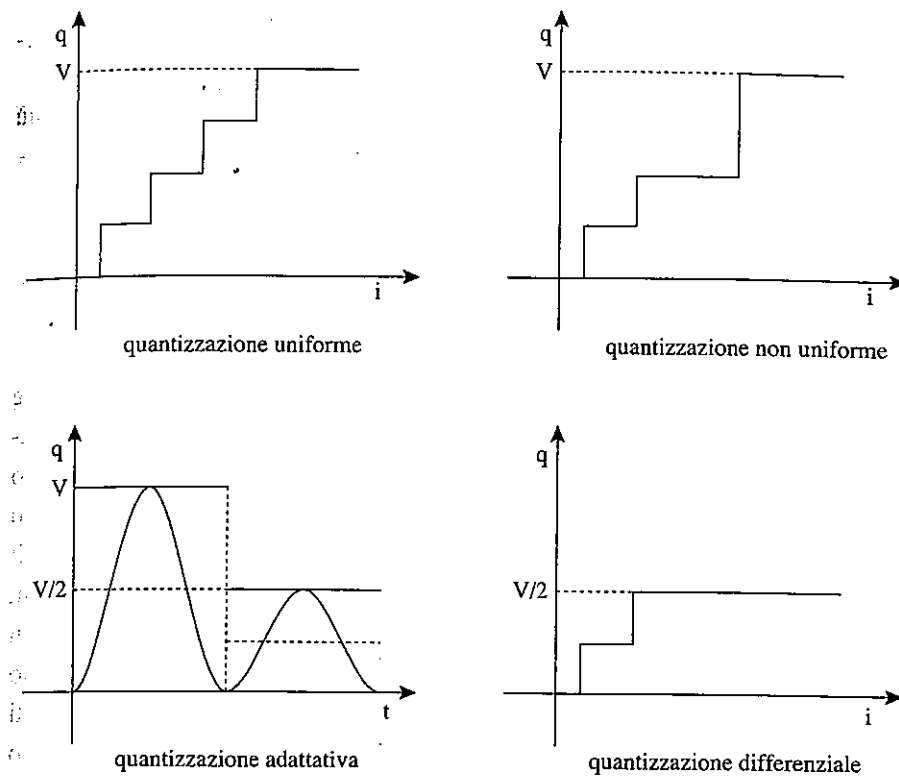



Fig. 2.3 - Riduzione del numero dei livelli del quantizzatore nelle differenti tecniche di codifica.

quanti in quelle bande nelle quali il rumore di quantizzazione risulterebbe mascherato dal segnale. Si hanno essenzialmente due tecniche di codifica in frequenza. La prima codifica, suddivide a blocchi lo spettro del segnale per mezzo di un banco di filtri passa banda. Dopo tale filtraggio, con le tecniche tipiche dei codificatori di forma d'onda, viene codificato il flusso continuo emesso da ciascun filtro (*codifica per sottobande*). La seconda codifica, suddivide in blocchi il flusso numerico, sottoponendo poi ciascun blocco ad una trasformazione per ottenere una rappresentazione continua in frequenza. I coefficienti ottenuti costituiscono la codifica del segnale (*codifica per trasfor-*

mate). Anche se l'implementazione risulta considerevolmente differente, queste due tecniche di codifica sono sostanzialmente equivalenti.

Nel seguito viene fornita un'analisi più dettagliata delle tecniche più diffuse per ciascuna delle codifiche descritte.

2.2 CONVERSIONE A/D - D/A E CODIFICA PCM LINEARE

2.2.1 Campionamento

Dato un segnale continuo, la sua rappresentazione numerica avviene mediante un processo di conversione Analogico/Digitale (A/D). La trasformazione inversa da digitale a continuo è detta conversione Digitale/Analogico (D/A). Essa consiste nel generare per ciascun codice il segnale analogico corrispondente. La conversione A/D si compone concettualmente di due fasi: un campionamento, che trasforma il segnale continuo in una serie di impulsi, ed una quantizzazione, che impone agli impulsi di assumere ampiezze in un insieme finito di valori. Gli impulsi generati dal campionamento si ottengono valutando il segnale in istanti non contigui di tempo. Tipicamente tali istanti sono equispaziati ed il tempo che intercorre tra due campioni adiacenti è detto periodo di campionamento (T). Il reciproco del periodo di campionamento è detta frequenza di campionamento (fs). La quantizzazione dei campioni su n livelli si ottiene fissando innanzitutto i valori massimo ed minimo del segnale d'uscita (estremi di saturazione) e poi suddividendo il campo delle ampiezze tra di essi comprese in n intervalli (quanti). La quantizzazione avviene poi associando ad un campione il livello corrispondente al valore medio dell'intervallo nel quale è compreso la sua ampiezza. Identificando ciascun livello con parole binarie di lunghezza pari a $R = \log_2(n)$ bit, si ottiene la codifica numerica voluta. Tale codifica è chiamata Pulse Code Modulation (PCM) lineare.

Per studiare le alterazioni subite dal segnale a seguito della conversione A/D e D/A, iniziamo ad analizzare gli effetti del campionamento. Come già detto, il campionamento periodico di un segnale continuo $x_c(t)$, fornisce una sequenza numerica x(n) costituita dai valori assunti dalla funzione a multipli del periodo di campionamento

$$x(n) = xc(n\,T); \supset n \qquad (2.2)$$

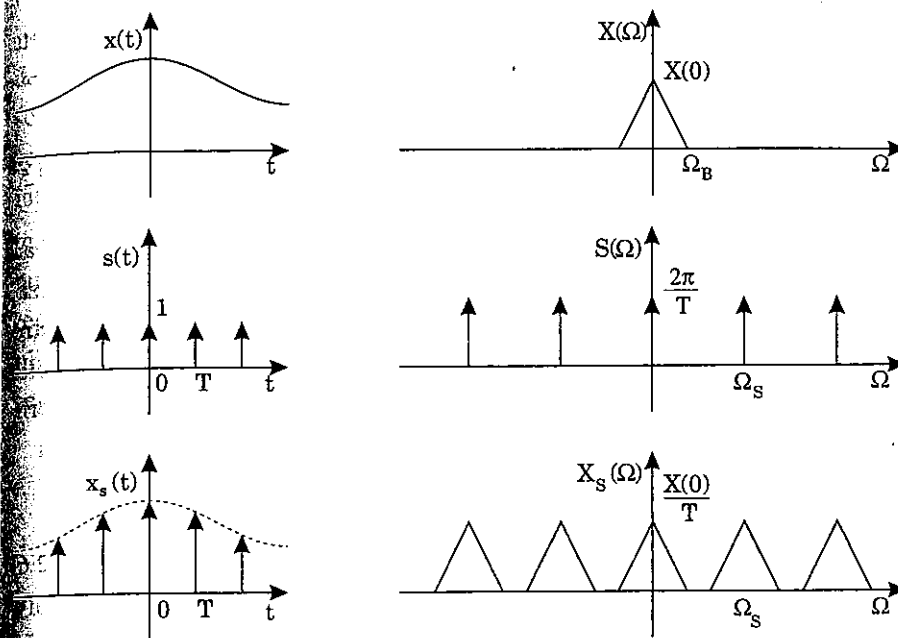



Fig. 2.4 - Effetti del campionamento sulla banda del segnale.

Il problema che si vuole affrontare è quello di determinare sotto quali condizioni tale trasformazione è reversibile, cioè, sotto quali condizioni è possibile ricostruire il segnale continuo dai suoi campioni. Si consideri una funzione continua xc(t) con una rappresentazione in frequenza $X_c(\Omega)$. Innanzitutto, è conveniente modellizzare il processo di campionamento come il risultato del prodotto del segnale $x_c(t)$ con una funzione campionatrice s(t) (fig. 2.4), costituita da un treno di δ posizionate negli istanti di campionamento

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{2.3}$$

Il risultato xs(t) di tale prodotto è un segnale PAM (Pulse Amplitude Modulation), costituito da un treno di impulsi, ciascuno dei quali ha ampiezza pari al valore che il segnale assume per t = nT

$$x_s(t) = x_c(t)\, s(t) = \sum_{n=-\infty}^{\infty} x_c(nT)\ \delta(t - nT) \tag{2.4}$$

È infine possibile trasformare il segnale PAM in una sequenza discreta x(n) tale che

$$x(n) = \lim_{\varepsilon \to 0} \int_{nT-\varepsilon}^{nT+\varepsilon} x_s(t)\, dt \qquad (2.5)$$

Per comprendere gli effetti del campionamento è vantaggioso passare ad una rappresentazione del segnale campionato nel dominio della frequenza. Dato che il segnale campionato è dato dal prodotto di due segnali, il suo spettro è ottenibile dalla convoluzione delle loro trasformate. Per il calcolo della trasformata S(f) della funzione di campionamento, conviene sviluppare la serie periodica di δ(t) in serie di Fourier. I coefficienti dello sviluppo sono pari a

$$c_k = \frac{1}{T}\int_{-T/2}^{T/2} \delta(t)\, e^{-jk\Omega_s t}\, dt = \frac{1}{T}\,; \quad \Omega_s = \frac{2\pi}{T} \qquad (2.6)$$

per cui la rappresentazione nel tempo del treno di delta è dato da

$$\sum_{n=-\infty}^{\infty} \delta(t-nT) = \frac{1}{T}\sum_{n=-\infty}^{\infty} e^{jn\Omega_s t} \qquad (2.7)$$

Calcolando la trasformata, si ha che la trasformata della costante 1/T è una δ(f), mentre la sommatoria di esponenziali porta a repliche della δ(f) traslate in frequenza, cioè

$$S(\Omega) = \frac{2\pi}{T}\sum_n \delta(\Omega - n\,\Omega_s) \qquad (2.8)$$

La rappresentazione in frequenza della funzione campionatrice, quindi, risulta essere ancora un treno di delta. Effettuando la convoluzione con lo spettro $X_c(\Omega)$ del segnale continuo, si ottiene la rappresentazione in frequenza $X_s(\Omega)$ del segnale campionato. Dato che la convoluzione di un segnale con la delta fornisce il segnale stesso, la convoluzione del segnale con un treno di delta fornisce una serie di sue repliche

$$X_s(\Omega) = \frac{1}{2\pi} X_c(\Omega) \otimes S(\Omega) = \frac{1}{T}\sum_k X_c(\Omega - k\Omega_s) \qquad (2.9)$$

Il campionamento, dunque, provoca la periodicizzazione in frequenza dello spettro del segnale originario, con un passo pari alla f_s. Per ricostruire il segnale continuo dai suoi campioni è necessario riottenere lo spettro del segnale continuo da quello del segnale campionato, cioè eliminare le repliche introdotte. Se la $x_c(t)$ è a banda limitata in un intervallo $[-\Omega_B, \Omega_B]$, l'eliminazione delle repliche è facilmente ottenibile tramite un filtraggio passa basso, purché queste non vadano a sovrapporsi alla $X_c(f)$. Dato che la posizione delle repliche è determinata dalla frequenza di campionamento, è necessario scegliere una frequenza di campionamento che sia almeno doppia della banda del segnale (teorema di Nyquist)

$$\Omega_s = \frac{2\pi}{T} = 2\pi f_s > 2\Omega_B \tag{2.10}$$

Per il filtro passa basso di ricostruzione, infine, la frequenza di taglio viene fissata alla metà della frequenza di campionamento (frequenza di Nyquist). Nel caso di segnali limitati in banda, quindi, è possibile ricostruire un segnale perfettamente identico all'originale, per cui il segnale $x_c(t)$ può essere univocamente determinato dal segnale $x_s(t)$.

Il filtro di ricostruzione è anche detto di interpolazione. Infatti, la ricostruzione può essere interpretata nel dominio del tempo come un'interpolazione tra campioni adiacenti (fig. 2.5). Per capire quale tipo di interpolazione venga utilizzata è necessario esprimere nel dominio del tempo l'effetto del prodotto in frequenza tra lo spettro del segnale campionato e la funzione di trasferimento del filtro. Se si considera un filtro passa basso ideale, la funzione di trasferimento del filtro è data da

$$H(\Omega) = \begin{cases} T & \text{per } |\Omega| < \pi/T \\ 0 & \text{altrove} \end{cases} \tag{2.11}$$

con una risposta impulsiva pari a

$$h_r(t) = \frac{\sin(t\pi/T)}{t\pi/T} = \text{sinc}\left(\frac{t\pi}{T}\right) \tag{2.12}$$

Il segnale ricostruito $x_r(t)$, dunque, si ottiene trasformando il prodotto dei due spettri in frequenza nella convoluzione nel dominio del tempo delle loro antitrasformate

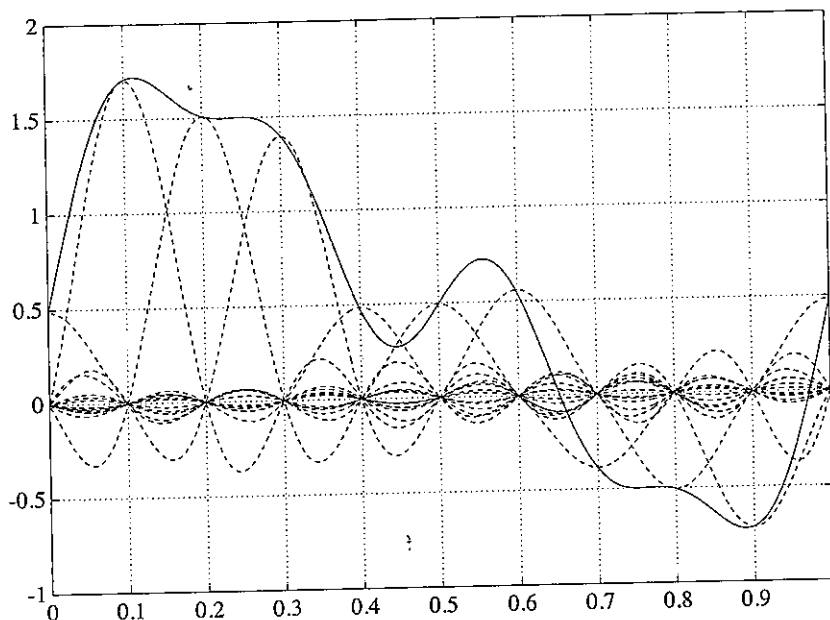



Fig. 2.5 - Ricostruzione di un segnale tramite interpolazione dei suoi campioni.

$$x_r(t) = x_s(t) \otimes h_r(t) = \sum_n x(n)\, h_r(t - nT) = \sum_n x(n) \frac{\sin[\pi(t-nT)/T]}{\pi(t-nT)/T} \qquad (2.13)$$

Analizzando questa relazione si nota che la $h_r(nT)$ vale 1 per n=0 ed è nulla per $n \neq 0$. L'ampiezza della $x_r(t)$ nei punti di campionamento, quindi, coincide con l'ampiezza dei campioni, mentre negli altri punti è ottenuta con una interpolazione tramite una funzione sinc(t).

È possibile dimostrare l'invertibilità del campionamento ideale anche procedendo in maniera differente. Si consideri la ripetizione periodica dello spettro di un segnale continuo a banda Ω_B limitata, con repliche poste a multipli pari della Ω_B

$$X_p(\Omega) = \sum_{k=-\infty}^{\infty} X_c(\Omega - 2k\Omega_B) \qquad (2.14)$$

Essendo la $X_p(\Omega)$ una funzione periodica, essa può essere sviluppata in serie di Fourier. Per un segnale funzione del tempo si ha

$$c_k = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\, e^{-j\frac{k2\pi}{T}t}\, dt\, ; \quad x(t) = \sum_{k=-\infty}^{\infty} c_k\, e^{j\frac{k2\pi}{T}T} \tag{2.15}$$

Nel nostro caso $t = \Omega$ e $T/2 = \Omega_B$. Inoltre, in un periodo la $X_p(\Omega)$ è descritta dalla $X_c(\Omega)$, per cui i coefficienti e lo sviluppo in serie si ottengono come

$$c_k = \frac{1}{2\Omega_B}\int_{-\Omega_B}^{\Omega_B} X_c(\Omega)\, e^{-j\frac{k\pi}{\Omega_B}\Omega}\, d\Omega\, ; \quad X_c(\Omega) = \sum_{k=-\infty}^{\infty} c_k\, e^{j\frac{k\pi}{\Omega_B}\Omega} \tag{2.16}$$

Considerando l'antitrasformata della $X_c(\Omega)$ (ricordando che questa è nulla al di fuori di Ω_B)

$$x_c(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X_c(\Omega)\, e^{j\Omega t}\, d\Omega = \frac{1}{2\pi}\int_{-\Omega_B}^{\Omega_B} X_c(\Omega)\, e^{j\Omega t}\, d\Omega \tag{2.17}$$

e sostituendo alla $X_c(\Omega)$ il suo sviluppo in serie, si ottiene

$$x_c(t) = \frac{1}{2\pi}\int_{-\Omega_B}^{\Omega_B} \sum_{k=-\infty}^{\infty} c_k\, e^{j\frac{k\pi}{\Omega_B}\Omega}\, e^{j\Omega t}\, d\Omega = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} c_k \int_{-\Omega_B}^{\Omega_B} e^{j\left(t+\frac{k\pi}{\Omega_B}\right)\Omega}\, d\Omega \tag{2.18}$$

Il calcolo dell'integrale fornisce

$$\int_{-\Omega_B}^{\Omega_B} e^{j\left(t+\frac{k\pi}{\Omega_B}\right)\Omega}\, d\Omega = \frac{1}{j\left(t+\frac{k\pi}{\Omega_B}\right)}\left[e^{j\left(t+\frac{k\pi}{\Omega_B}\right)\Omega}\right]_{-\Omega_B}^{\Omega_B} = \frac{e^{j\left(t+\frac{k\pi}{\Omega_B}\right)\Omega_B} - e^{-j\left(t+\frac{k\pi}{\Omega_B}\right)\Omega_B}}{j\left(t+\frac{k\pi}{\Omega_B}\right)}$$

$$= \frac{2\sin\left[\Omega_B\left(t+\frac{k\pi}{\Omega_B}\right)\right]}{t+\frac{k\pi}{\Omega_B}} = 2\,\Omega_B\, \mathrm{sinc}\left[\Omega_B\left(t+\frac{k\pi}{\Omega_B}\right)\right] \tag{2.19}$$

quindi

$$x_c(t) = \frac{\Omega_B}{\pi}\sum_{k=-\infty}^{\infty} c_k\, \mathrm{sinc}\left[\Omega_B\left(t+\frac{k\pi}{\Omega_B}\right)\right] \tag{2.20}$$

La ricostruzione del segnale avviene, quindi, tramite delle sinc(t) posizionate a multipli del periodo di campionamento. Dato che la sinc(nT) vale 1 per n=0 ed è nulla per $n \neq 0$, il valore dei coefficienti possono essere determinati imponendo che le ampiezze fornite dalla serie coincidano negli istanti di campionamento con quelle della xc(t). In tal modo si ottiene

$$c_k = \frac{\pi}{\Omega_B} x_c\left(t + \frac{k\pi}{\Omega_B}\right) = \frac{\pi}{\Omega_B} x_c\left(t - \frac{n\pi}{\Omega_B}\right) = T\ x_s(nT)\ ;$$

$$\begin{cases} k \in (-\infty, +\infty) \\ n = -k \end{cases} \qquad (2.21)$$

I coefficienti della serie dello spettro periodicizzato, quindi, sono dati da campioni scalati del segnale. Rovesciando il discorso, il segnale campionato ha una rappresentazione in frequenza data dalla periodicizzazione del suo spettro.

Per segnali a banda limitata, dunque, affinché il campionamento sia invertibile è sufficiente rispettare il teorema di Nyquist. Se ciò non avviene le repliche dovute al campionamento si sovrappongono allo spettro del segnale originario (aliasing), impedendone una successiva separazione. In pratica, nessun segnale è limitato in banda. La principale causa è la presenza inevitabile del rumore: in tal caso l'aliasing somma al segnale infinite repliche dello spettro del rumore, degradandone la qualità. Inoltre, accade che sullo stesso mezzo possano viaggiare oltre al segnale di interesse, altri segnali multiplexati in frequenza tramite modulazione (es.: filodiffusione): campionando con una frequenza legata al solo segnale non modulato, la separazione in frequenza viene distrutta a causa delle repliche del segnale modulato generate su tutto lo spettro. Per poter fissare la frequenza di campionamento, quindi, è necessario limitare forzatamente il segnale prima del suo campionamento. Ciò si ottiene tramite un filtro passa basso (anti-aliasing) con frequenza di taglio pari alla metà della frequenza di campionamento. Si pone l'accento sul fatto che, anche se le caratteristiche del filtro anti-aliasing sono analoghe a quelle del filtro di interpolazione le sue funzioni sono completamente differenti.

Nel caso di segnale telefonico, il campionamento impone che la banda del segnale venga limitata a 3.4 kHz e la frequenza di campionamento è fissata a 8 kHz. Per un segnale audio HiFi, avente banda di 20 kHz, la frequenza di campio-

namento è di 44.1 kHz nel Compact Disk (CD) e di 48 kHz nel Digital Audio Tape (DAT) per garantire un certo intervallo di transizione ai filtri anti aliasing e di ricostruzione. Nel caso delle trasmissioni radio (Digital Audio Broadcasting: DAB) la frequenza di campionamento è di 32 kHz per una banda di 15 kHz.

Nella trattazione precedente si è considerato un campionamento ideale, ottenuto tramite un treno di δ. In realtà, è necessario tener presente il funzionamento a tenuta (hold) dei convertitori D/A che fa si che l'uscita non sia composta da impulsi, ma il livello di ciascun campione è mantenuto costante all'interno dell'intervallo di campionamento, per essere modificato all'arrivo del campione successivo (fig. 2.6). Analiticamente ciò è ottenibile eseguendo la convoluzione dei campioni ideali con una rect(t):

$$x_s(t) = [\, x_c(t)\, s(t)\,] \otimes \mathrm{rect}(t) = \left[x_c(t) \sum_{n=-\infty}^{\infty} \delta(t-nT) \right] \otimes \mathrm{rect}(t) \tag{2.22}$$

In frequenza ciò si riflette nel prodotto dello spettro del segnale campionato con la trasformata della rect(t), che è una sinc(f)

$$X_s(f) = \left[X_c(f) \otimes \delta\left(f - \frac{n}{T}\right)\right] \frac{\sin(\pi f)}{\pi f} \tag{2.23}$$

In definitiva, lo spettro di un segnale ottenuto dal convertitore D/A è dato dalla ripetizione periodica dello spettro del segnale originario, come per il campionamento ideale, ma con la differenza che l'ampiezza non è costante, ma decresce in frequenza con un inviluppo dato da una sinc(f).

Nella presente trattazione sia il filtro anti-aliasing che quello di interpolazione sono stati considerati ideali. L'utilizzo di filtri con funzione di trasferimento non ideale (distorsioni in banda, attenuazione finita fuori banda, ecc.) porta a degradazioni delle prestazioni, che, però, non vengono qui ulteriormente approfondite.

2.2.2 Quantizzazione lineare

Un campione x(n) del segnale è una misura dell'ampiezza del segnale stesso che può assumere qualsiasi valore nell'intervallo (−∞, +∞) ed è, quindi, esprimibile come numero reale. Al fine di permettere la rappresentazione del

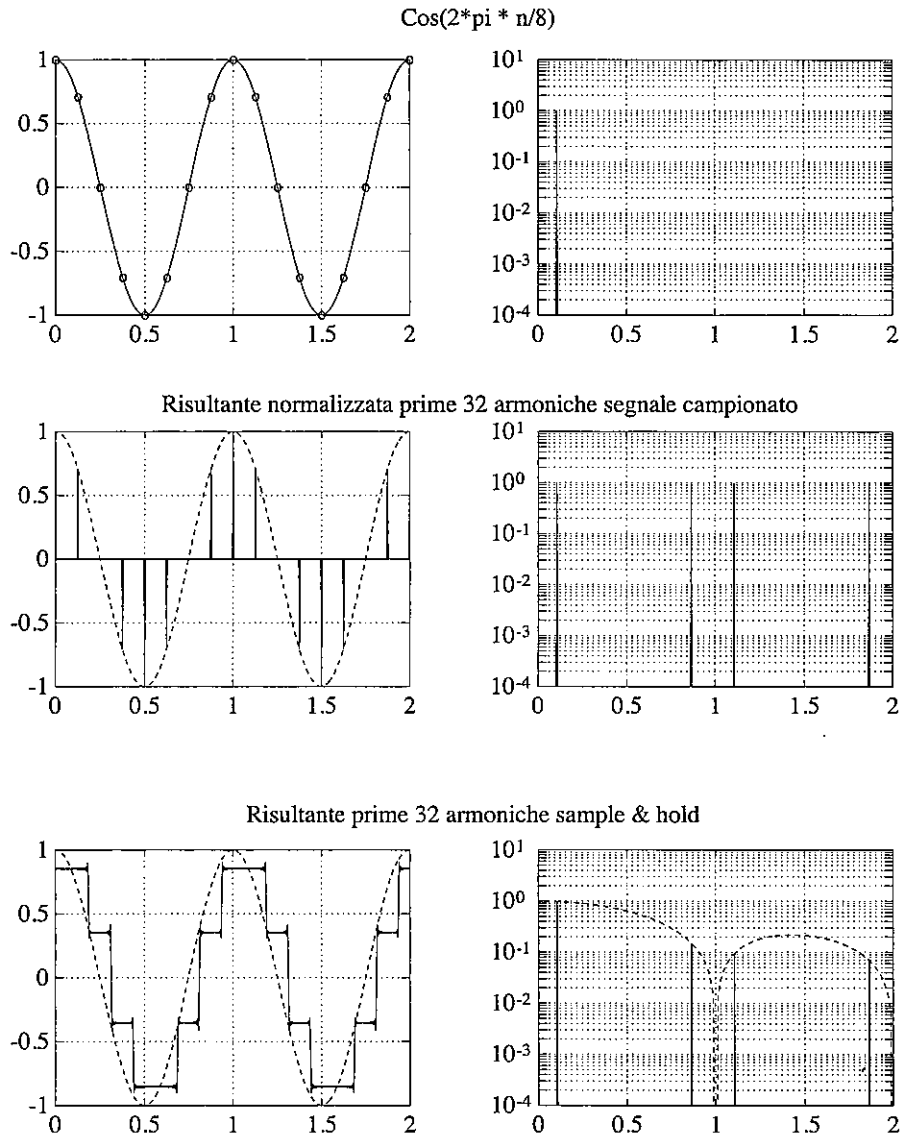



Fig. 2.6 - Forma d'onda e spettri per sample & hold.

segnale campionato tramite parole binarie con precisione finita, il segnale analogico deve essere sottoposto a quantizzazione. Il più semplice algoritmo di quantizzazione è la quantizzazione uniforme. Con tale quantizzazione è necessario, innanzitutto, fissare dei livelli minimi e massimi di ampiezze

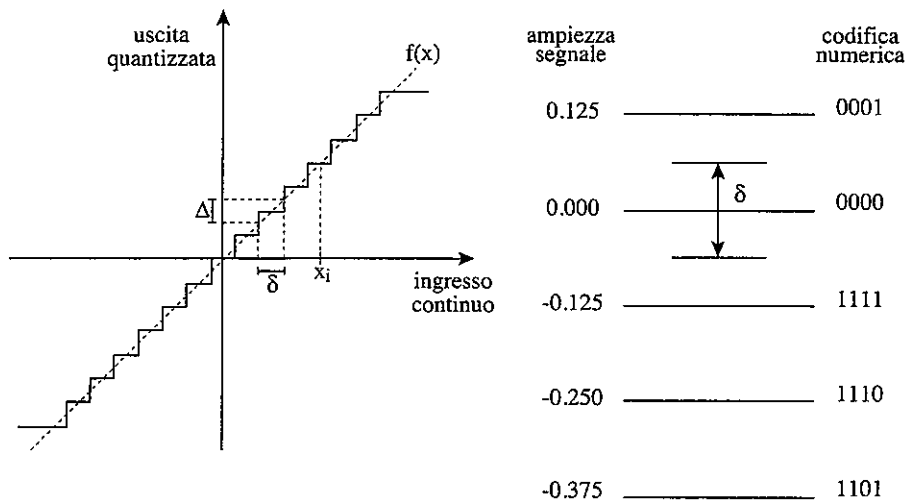



Fig. 2.7 - Quantizzazione uniforme.

ammissibili della grandezza d'ingresso. Per segnali a valor medio nullo, tale intervallo è tipicamente simmetrico rispetto allo zero, per cui le ampiezze ammissibili sono quelle contenute tra due estremi +/- V, detti estremi di saturazione. L'intervallo della ampiezze all'interno degli estremi di saturazione è poi suddiviso in sotto intervalli δ_{x_i}, detti quanti (fig. 2.7).

La quantizzazione consiste nell'individuare il quanto entro il quale cade l'ampiezza del campione d'ingresso e nel sostituire all'informazione dell'ampiezza del campione stesso l'indice di tale quanto, opportunamente codificato. Tale numero può essere poi trasmesso o memorizzato per permettere la ricostruzione del segnale. Utilizzando una codifica binaria di lunghezza fissa, con un numero di livelli pari a $n = 2R$, l'ampiezza del quanto è pari a $\Delta = 2V/n$ ed il codice prodotto è formato da interi rappresentabili su R bit. A tutti i valori di ampiezza al di fuori degli estremi di saturazione viene associato il codice dell'ultimo quanto utile.

In fase di ricostruzione del segnale a partire dai suoi campioni quantizzati, è necessario associare un'ampiezza a ciascun quanto. Tale ampiezza è normalmente scelta pari al valor medio del quanto stesso. Dato che ad intervalli di ampiezze dell'ingresso si fa coincidere un solo valore dell'uscita, la quantizzazione rappresenta una compressione con riduzione di entropia della sorgente e la distorsione che ne deriva è irreversibile. Se applicata all'uscita di un

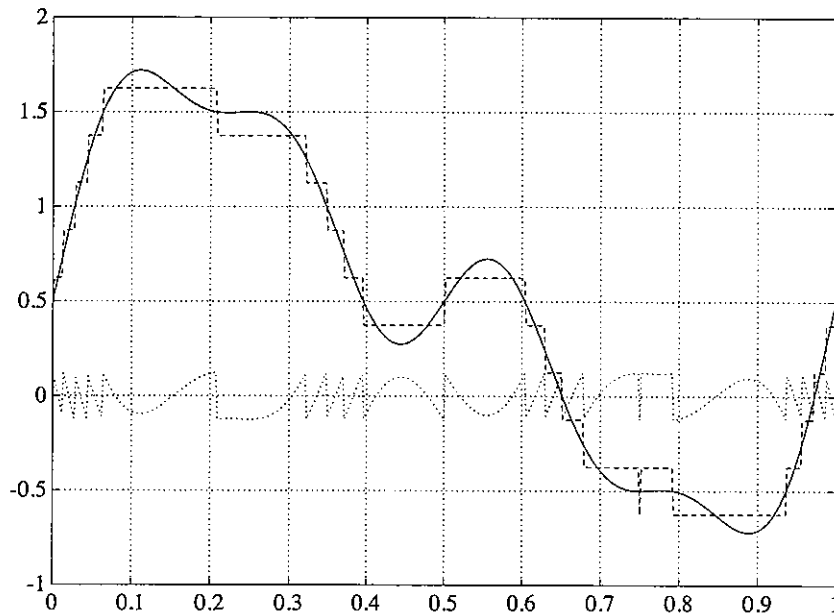



Fig. 2.8 - Errore di quantizzazione.

campionatore ideale, dalla sequenza $\hat{x}(n)$ ottenuta dalla quantizzazione non è quindi possibile riottenere la sequenza x(n) generata dal campionamento, per cui non è possibile ricostruire esattamente il segnale xc(t).

Guardandola nel suo complesso, la quantizzazione è schematizzabile come una trasformazione non lineare che mappa intervalli di ampiezza costante δ_{x_i} del segnale in ingresso in un insieme finito di ampiezze. In un quantizzatore uniforme, queste risultano tra loro equispaziate e la distanza tra due ampiezze di uscita adiacenti coincide con l'ampiezza Δ del quanto d'ingresso. La legge di corrispondenza f(x) tra il valor medio dei quanti d'ingresso e la corrispondente ampiezza d'uscita è detta *caratteristica di quantizzazione*. Nel caso di quantizzazione uniforme, la caratteristica è costituita dalla bisettrice del primo e terzo quadrante (fig. 2.7).

La distorsione introdotta dalla quantizzazione è modellizzabile con la somma del segnale d'ingresso con un secondo segnale, detto rumore di quantizzazione, dato dalla differenza tra l'ingresso e l'uscita. Se si paragona la quantizzazione alla trasformazione di un numero reale in un intero, l'errore di quantizzazione rappresenta l'analogo dell'errore di arrotondamento. La poten-

za del rumore di quantizzazione è il parametro che fissa il limite massimo del rapporto segnale/rumore ottenibile in decodifica (fig. 2.8). Nel seguito si vogliono individuare i parametri che influenzano l'entità del rumore di quantizzazione, per fissarne l'ampiezza in funzione delle caratteristiche del segnale.

Abbandonando, per generalità, l'ipotesi di caratteristica lineare, è possibile, innanzitutto, calcolare la dimensione dell'intervallo di ampiezze i-esimo del segnale continuo d'ingresso δ_{x_i} che corrisponde ad un quanto costante Δ d'uscita. Approssimando la caratteristica in ciascun intervallo con la sua tangente nel punto medio xi si ottiene

$$\delta_{x_i} \approx \frac{\Delta}{f'(x_i)} = \frac{2V}{n\, f'(x_i)} = \frac{2V}{2^R f'(x_i)} \qquad (2.24)$$

In ciascun intervallo è poi possibile calcolare l'errore quadratico medio tra il segnale e la sua versione quantizzata (cioè l'energia dell'errore di quantizzazione), in funzione della distribuzione di probabilità p(x) delle ampiezze del segnale

$$e_q^2 = E\left\{(x - x_i)^2\right\} = \int_{x_i - \frac{\delta_{x_i}}{2}}^{x_i + \frac{\delta_{x_i}}{2}} (x - x_i)^2\, p(x)\, dx \qquad (2.25)$$

Se il numero di livelli è elevato, la probabilità all'interno del quanto si può ritenere costante e pari a quella del valor medio $p(x_i)$, per cui

$$e_q^2 \approx \int_{x_i - \frac{\delta_{x_i}}{2}}^{x_i + \frac{\delta_{x_i}}{2}} (x - x_i)^2\, p(x_i)\, dx = p(x_i) \int_{x_i - \frac{\delta_{x_i}}{2}}^{x_i + \frac{\delta_{x_i}}{2}} (x - x_i)^2\, dx = p(x_i) \left[\frac{(x - x_i)^3}{3} \right]_{x_i - \frac{\delta_{x_i}}{2}}^{x_i + \frac{\delta_{x_i}}{2}}$$

$$e_q^2 = \frac{\delta_{x_i}^3}{12}\, p(x_i) = \frac{\delta_{x_i}^2}{12}\, P(x_i) = \frac{V^2}{3\, n^2 f'^2(x_i)}\, P(x_i) \qquad (2.26)$$

dove $P(x_i) = p(x_i)\ \delta_{x_i}$ rappresenta la probabilità che il segnale sia compreso nel quanto. Nell'ipotesi di p(x) ed f(x) simmetrici rispetto all'origine e considerando intervalli infinitesimi dx, la $P(x_i)$ è esprimibile come p(x) dx e l'errore quadratico totale è pari a

$$e_q^2 = \int_{-\infty}^{\infty} \frac{V^2}{3\,n^2\,f'^2(x)}\,p(x)\,dx = \frac{2V^2}{3n^2}\int_0^V \frac{p(x)}{f'^2(x)}\,dx + 2\int_V^{\infty}(x-V)^2\,p(x)\,dx$$

$$e_q^2 = e_g^2 + e_s^2 \tag{2.27}$$

Il primo termine dell'espressione dell'errore tiene conto della distorsione del segnale dovuto alla quantizzazione (rumore granulare di quantizzazione o, brevemente, rumore di quantizzazione), mentre il secondo termine riguarda la distorsione dovuta alla saturazione (rumore di sovraccarico).

Considerando il solo rumore di sovraccarico, la sua potenza

$$e_s^2 = 2\int_V^{\infty}(x-V)^2\,p(x)\,dx \tag{2.28}$$

dipende dal valore della V. Questo deve essere fissato in funzione della distribuzione di probabilità delle ampiezze del segnale p(x) in modo tale che il rumore introdotto dalla saturazione possa essere considerato trascurabile secondo qualche criterio. Se si considera, ad esempio, un segnale sinusoidale $x = A\sin(\vartheta)$, con $A > V$ e $\theta \in (0, 2\pi)$, e considerata uniforme la distribuzione di probabilità dell'argomento ϑ ($p(\vartheta) = 1/2\pi$), la distribuzione di probabilità della x è pari a [Pap84]

$$p(x) = 2\,p(\theta)\frac{d\theta}{dx} = \frac{1}{\pi\sqrt{A^2 - x^2}}\ ; \quad |x| < A \tag{2.29}$$

La potenza del rumore di saturazione è, quindi, pari a

$$e_s^2 = 2\int_V^{\infty}(V-x)^2\,p(x)\,dx = \frac{2}{\pi}\int_V^{A}\frac{(V-x)^2}{\sqrt{A^2-x^2}}\,dx \tag{2.30}$$

Con la sostituzione

$$t = \sin^{-1}\frac{x}{A};\ dt = \frac{dx}{\sqrt{A^2 - x^2}} \tag{2.31}$$

si ottiene

$$e_s^2 = \frac{2}{\pi}\int_V^A \frac{(V-x)^2}{\sqrt{A^2-x^2}}\,dx = \frac{2}{\pi}\int_\varphi^{\frac{\pi}{2}} [\,V - A\sin(t)\,]^2\,dt$$

$$= \frac{2}{\pi}\int_\varphi^{\frac{\pi}{2}} \left[\,V^2 + A^2\sin(t)^2 - 2AV\sin(t)\,\right]\,dt$$

$$= \frac{2}{\pi}\left\{ V^2 t + A^2\left[-\frac{1}{4}\sin(2t) + \frac{1}{2}t\right] + 2AV\cos(t)\right\}_\varphi^{\frac{\pi}{2}}$$

$$= \frac{2}{\pi}\left\{\left(V^2 + \frac{A^2}{2}\right)t - \frac{A^2}{4}\sin(2t) + 2AV\cos(t)\right\}_\varphi^{\frac{\pi}{2}} \tag{2.32}$$

cioè

$$e_s^2 = \frac{2}{\pi}\left\{\left(V^2 + \frac{A^2}{2}\right)\left(\frac{\pi}{2} - \varphi\right) + \frac{A^2}{4}\sin(2\varphi) - 2AV\cos(\varphi)\right\} \tag{2.33}$$

Nota la potenza del rumore di sovraccarico e noto che per un segnale sinusoidale $x_{eff} = A / \sqrt{2}$, è possibile calcolare il rapporto segnale/rumore per tale componente, che, espresso in dB, è dato da

$$\frac{S}{N} = 10\log\frac{x_{eff}^2}{e_s^2} = 20\log\frac{x_{eff}}{e_s} \tag{2.34}$$

Il rapporto segnale rumore viene generalmente espresso in funzione del valore normalizzato di xeff rispetto al livello di pieno carico sinusoidale xeffP

$$10\log\frac{x_{eff}^2}{x_{effP}^2} = 20\log\frac{x_{eff}}{x_{effP}} \tag{2.35}$$

dove xeffP è definito come il valore efficace della sinusoide che ha come valore di picco la tensione di saturazione V, cioè xeffP = V / $\sqrt{2}$. Ovviamente, nel caso

di segnali sinusoidali il rapporto S/N va all'infinito per $A \leq V$, per cui la scelta più ovvia per il limite di saturazione è $V \geq A$.

Passando a segnali non deterministici, una distribuzione di particolare interesse è quella gaussiana

$$p(x) = \sqrt{\frac{\beta}{\pi}}\, e^{-\beta x^2} = \frac{1}{\sigma_x \sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_x^2}x^2}\ ; \quad x \in (-\infty, +\infty) \tag{2.36}$$

dove $\beta = 1 / (2\,\sigma_x^2) = 1 / (2\,x_{eff}^2)$, in quanto corrispondente alla distribuzione a breve termine (intervalli di circa 10 ms) delle ampiezze per il segnale telefonico [Jay84]. L'espressione per l'errore di sovraccarico che si ottiene nel caso di distribuzione gaussiana è data da

$$e_s^2 = 2\sqrt{\frac{\beta}{\pi}} \int_V^\infty (x-V)^2\, e^{-\beta x^2}\, dx$$

$$= 2\sqrt{\frac{\beta}{\pi}} \left\{ \int_V^\infty x^2 e^{-\beta x^2}\, dx \ - \ 2V \int_V^\infty x\, e^{-\beta x^2}\, dx \ + \ V^2 \int_V^\infty e^{-\beta x^2}\, dx \right\} \tag{2.37}$$

Procedendo al calcolo dei singoli integrali si ottiene

$$\int_V^\infty e^{-\beta x^2}\, dx = \frac{1}{2}\sqrt{\frac{\pi}{\beta}}\ [\,1 - \mathrm{erf}(\sqrt{\beta}\, V)\,]\ ; \quad \text{(per definizione)}$$

$$\int_V^\infty x\, e^{-\beta x^2}\, dx = \frac{1}{2\beta} e^{-\beta V^2}\ ; \quad \text{(tramite sostituzione } t = \beta x2\text{)}$$

$$\int_V^\infty x^2 e^{-\beta x^2}\, dx = \frac{V}{2\beta} e^{-\beta V^2} + \frac{\sqrt{\pi}}{4\sqrt{\beta^3}} [\,1 - \mathrm{erf}(\sqrt{\beta}\, V)\,]\ ;$$

(2.38)

per cui

$$e_s^2 = 2\sqrt{\frac{\beta}{\pi}} \left\{ \frac{\sqrt{\pi}}{4\sqrt{\beta^3}}\ (1 + 2\beta V^2)\ [\,1 - \mathrm{erf}(\sqrt{\beta}\, V)\,] \ - \ \frac{V}{2\beta} e^{-\beta V^2} \right\}$$

$$= \sigma_x \sqrt{\frac{2}{\pi}} \left\{ \sigma_x \sqrt{\frac{\pi}{2}} \left(1 + \frac{V^2}{\sigma_x^2}\right) \left[1 - \mathrm{erf}\left(\frac{V}{\sigma_x\sqrt{2}}\right)\right] \ - \ V\, e^{-\frac{1}{2}\frac{V^2}{\sigma_x^2}} \right\} \tag{2.39}$$

Sempre per il segnale telefonico, una buona approssimazione della distribuzione delle ampiezze a lungo termine (intervalli dell'ordine del secondo) è data dalla distribuzione gamma [Jay84], che ha densità di probabilità

$$p(x) = \begin{cases} \dfrac{x^{\varepsilon-1}}{\delta^{\varepsilon}\,\Gamma(\varepsilon)}\, e^{-\frac{x}{\delta}}; & x > 0 \\ 0; & x \leq 0 \end{cases} \rightarrow \frac{\sqrt[4]{3}}{\sqrt{8\,\pi\,\sigma_x\,|x|}}\; e^{-\frac{\sqrt{3}}{2\sigma_x}|x|} \tag{2.40}$$

Un'ulteriore utile approssimazione di questa distribuzione è quella esponenziale (o di Laplace), che ha densità di probabilità

$$p(x) = \frac{\alpha}{2}\, e^{-\alpha|x|} = \frac{\sqrt{2}}{2\,\sigma_x}\, e^{-\frac{\sqrt{2}}{\sigma_x}|x|}; \quad x \in (-\infty, +\infty) \tag{2.41}$$

dove $\alpha = \sqrt{2}\,/\,\sigma_x$. In tal caso il calcolo dell'errore di sovraccarico fornisce

$$e_s^2 = 2\int_V^{\infty} (x - V)^2\, p(x)\, dx = \alpha \int_V^{\infty} \left(x^2 - 2Vx + V^2\right) e^{-\alpha x}\, dx$$

$$= -\left\{\left[x^2 - 2\left(\frac{1}{\alpha} + V\right)x + \frac{2}{\alpha^2} + \frac{2V}{\alpha} + V^2\right] e^{-\alpha x}\right\}_V^{\infty} = \frac{2}{\alpha^2}\, e^{-\alpha V}$$

$$e_s^2 = \sigma_x^2\, e^{-\frac{\sqrt{2}}{\sigma_x}V} \tag{2.42}$$

Analizzando l'andamento del rumore di sovraccarico per queste tre distribuzioni (fig. 2.9), si nota come l'effetto della saturazione sia tanto maggiore quanto maggiore è il fattore di cresta del segnale, cioè del rapporto tra valore di picco e valore efficace (il valore di picco è definito come il valore del segnale che non viene superato con probabilità del 99 %). Confrontando, ad esempio, un segnale con distribuzione gaussiana (la cui distribuzione decresce proporzionalmente a e^{-x^2}) con un segnale con distribuzione esponenziale (la cui distribuzione decresce proporzionalmente a e^{-x}), si nota come, a parità di valore efficace, il secondo risenta maggiormente degli effetti della saturazione a causa della maggiore estensione delle code della funzione densità di probabilità.

Riassumendo, fissato il limite di saturazione, la potenza della relativa componente dell'errore di quantizzazione dipende dalla funzione densità di

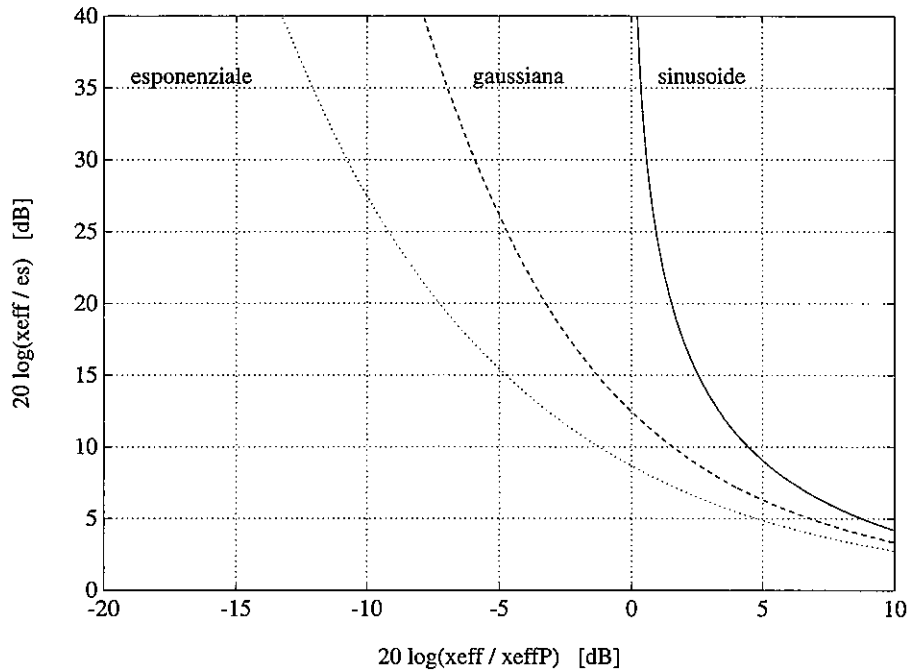



Fig. 2.9 - Rapporto S/N di sovraccarico per differenti distribuzioni.

probabilità del segnale. Nel caso di segnale telefonico gli estremi V del quantizzatore sono standardizzati a 3.14 dBm0 [ITU-T G.711].

Se i limiti del quantizzatore sono stati scelti opportunamente, la componente del rumore dovuta al sovraccarico si può ritenere trascurabile, per cui il rumore di quantizzazione è dato dalla sola componente granulare. È intuitivo pensare che, dato che il rumore è funzione dello scostamento tra la funzione continua e la sua versione quantizzata e dato che questo scostamento dipende dall'ampiezza dell'intervallo di quantizzazione, il rumore di quantizzazione dipenda dal numero dei livelli del quantizzatore, quindi dal numero di bit utilizzati per il codice. Infatti, data l'espressione della componente granulare

$$e_g^2 = \frac{2V^2}{3n^2} \int_0^V \frac{p(x)}{f'^2(x)}\, dx \tag{2.43}$$

e considerando il caso di quantizzazione uniforme ($f'(x_i) = 1 \rightarrow \delta_{x_i} = \Delta$) e distribuzione di probabilità è costante, risulta

$$2\int_0^V \frac{p(x)}{f'^2(x)}\,dx \cong 2\int_0^V p(x)\,dx \cong 1 \tag{2.44}$$

per cui

$$e_g^2 = \frac{\Delta^2}{12} = \frac{V^2}{3n^2} = \frac{V^2}{3}2^{-2R} = \varepsilon_x^2\sigma_x^2\,2^{-2R} \tag{2.45}$$

dove ε_x^2 è una costante (> 1) che lega gli estremi di saturazione alla varianza del segnale. Conformemente a quanto ci si aspettava, dunque, il rumore di quantizzazione è inversamente proporzionale al numero di bit utilizzati nella codifica. Per semplificare ulteriormente tale relazione, è conveniente esprimerla in unità logaritmiche, ottenendo

$$10\log e_g^2 = 10\log V^2 - 10\log 3 - 10\log 2^{2R} = 20\log V - 4.77 - 20\,R\log 2 \tag{2.46}$$

Trascurando i termini costanti che appaiono nell'equazione, si ottiene

$$e_g^2 \approx -6R\ \text{dBm0} \tag{2.47}$$

Questa è la cercata relazione che lega il rumore di quantizzazione al numero di bit utilizzati nella codifica. Essa mostra che il rumore si riduce di 6 dB per ogni bit utilizzato nel quantizzatore. A questo punto è possibile fissare il numero di bit utilizzati nella quantizzazione in funzione delle caratteristiche del segnale (fig. 2.10). Il parametro da valutare è la dinamica, cioè il rapporto tra il massimo ed il minimo livello utile del segnale. Infatti, scelto il livello di saturazione del quantizzatore pari al massimo livello del segnale, il minimo livello utile del segnale stesso deve risultare superiore al livello del rumore di quantizzazione. Ponendo pari a 0 dB il livello massimo del segnale, il minimo è posto ad un livello numericamente pari alla dinamica. Per fissare l'errore di quantizzazione al di sotto del livello minimo tramite la relazione approssimata trovata, è necessario adottare nel quantizzatore un numero di bit maggiore del valore che si ottiene dividendo la dinamica per 6. Per il segnale telefonico, ad esempio, caratterizzato da una dinamica maggiore di 50 dB, è necessario adottare una quantizzazione lineare con un numero di bit maggiore di 10. Passando alle ampiezze, porre la potenza del rumore di quantizzazione al di

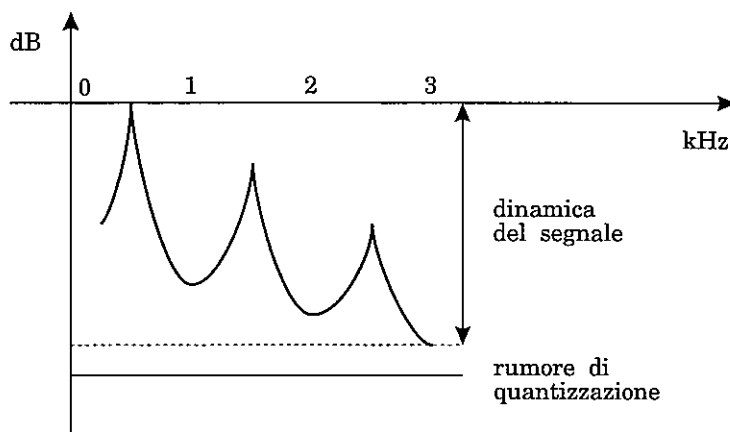



Fig. 2.10 - Dinamica del segnale e rumore di quantizzazione.

sotto di quella del più debole segnale utile vuol dire scegliere un'ampiezza del quanto inferiore a quella di tale segnale, permettendone la rappresentabilità.

Mantenendo l'ipotesi di quantizzazione uniforme, ma abbandonando l'ipotesi di distribuzione uniforme di probabilità delle ampiezze del segnale, l'espressione del rumore granulare va modificata inserendovi quella della probabilità. Considerando una distribuzione esponenziale si ottiene

$$e_g^2 = \frac{2V^2}{3n^2} \int_0^V p(x)\, dx = \frac{2V^2}{3n^2} \frac{\alpha}{2} \int_0^V e^{-\alpha x}\, dx = \frac{V^2}{3n^2} \left(1 - e^{-\alpha V}\right) \tag{2.48}$$

Se si confrontano i risultati ottenuti con tale espressione con quelli dati dall'approssimazione di distribuzione uniforme, non si notano differenze apprezzabili (dell'ordine di 10^{-3} per $x_{eff} / x_{effP} = 0$ dB ed inferiori per ampiezze minori).

Per quanto riguarda l'andamento del rapporto segnale rumore (fig. 2.11), si nota come la componente granulare del rumore di quantizzazione appaia solo in presenza del segnale (a meno dell'idle channel noise descritto nel seguito) e risulti indipendente dall'ampiezza del segnale stesso. Di conseguenza, il rapporto segnale rumore

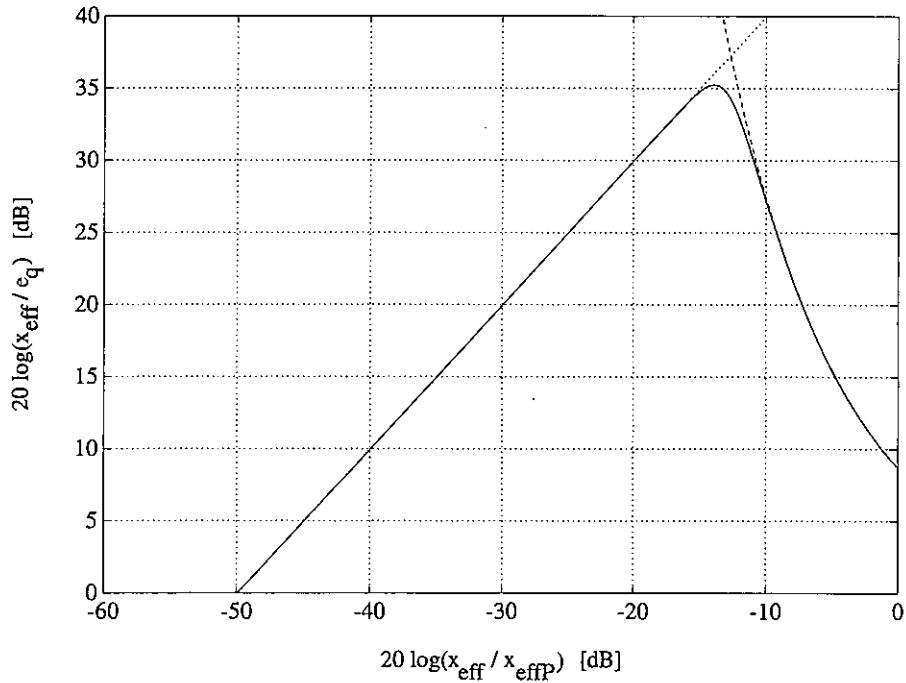



Fig. 2.11 - Errore di quantizzazione per distribuzione esponenziale.

$$SNR = \frac{x_{eff}^2}{e_g^2} = \frac{12}{\Delta^2} x_{eff}^2 \tag{2.49}$$

migliora linearmente con l'aumentare della potenza del segnale. Nel caso più generale in cui la componente granulare è sommata alla componente dovuta al sovraccarico, il rapporto segnale rumore cresce linearmente fino a che non si entra in saturazione, dopo di che si ha un brusco peggioramento. Tale miglioramento con l'ampiezza del segnale non risulta vantaggioso dal punto di vista della codifica, dato che non si sfrutta in tal modo il potere di mascheramento che il segnale ha sul rumore. Tecniche di quantizzazione non uniforme, esposte nel seguito, intervengono in tal senso.

2.2.3 Quantizzazione uniforme ottima

Si vuole determinare la struttura di un quantizzatore in grado di fornire il minimo valore di distorsione e_q^2 per un dato numero di bit per campione. Per un quantizzatore uniforme, ciò si traduce nella determinazione del suo passo di quantizzazione dato che, fissato R, ne viene contemporaneamente fissato l'altro suo parametro caratteristico, che è il livello di saturazione. È evidente che se una diminuzione del passo di quantizzazione riduce la componente granulare del rumore, dall'altra ne aumenta la componente di sovraccarico e viceversa, per cui il passo di quantizzazione ottimo sarà tale da minimizzare contemporaneamente tali componenti. Indicando con y l'uscita del quantizzatore, la distorsione media provocata dalla quantizzazione, utilizzando come sua misura lo scarto quadratico medio, è pari a

$$D = \int_{-\infty}^{\infty} (y-x)^2 \, p(x) \, dx \tag{2.50}$$

che per un quantizzatore uniforme diventa

$$D = 2 \sum_{k=1}^{L/2-1} \int_{(k-1)\Delta}^{k\Delta} \left[\frac{(2k-1)\Delta}{2} - x\right]^2 p(x) \, dx + 2 \int_{(L/2-1)\Delta}^{\infty} \left[\frac{(L-1)\Delta}{2} - x\right]^2 p(x) \, dx \tag{2.51}$$

Il minimo di tale funzione si ottiene annullando la derivata rispetto alla Δ come

$$\sum_{k=1}^{L/2-1} (2k-1) \int_{(k-1)\Delta}^{k\Delta} \left[\frac{(2k-1)\,\Delta}{2} - x\right]^2 p(x) \, dx + 2\,(L-1) \int_{(L/2-1)\Delta}^{\infty} \left[\frac{(L-1)\,\Delta}{2} - x\right] p(x) \, dx = 0 \tag{2.52}$$

La soluzione a questo problema è ricavabile per via numerica e viene riportata in tabella per le distribuzioni di probabilità uniforme, gaussiana, laplaciana e gamma [Jay84]:

R	Δ / σ_X				SNR max (dB)			
	U	G	L	Γ	U	G	L	Γ
1	1.7320	1.5956	1.4142	1.1547	6.02	4.40	3.01	1.76
2	0.8660	0.9957	1.0874	1.0660	12.04	9.25	7.07	4.95
3	0.4330	0.5860	0.7309	0.7957	18.06	14.27	11.44	8.78
4	0.2165	0.3352	0.4610	0.5400	24.08	19.38	15.96	13.00
5	0.1083	0.1881	0.2800	0.3459	30.10	24.57	20.60	17.49
6	0.0541	0.1041	0.1657	0.2130	36.12	29.83	25.36	22.16
7	0.0271	0.0569	0.0961	0.1273	42.14	35.13	30.23	26.99
8	0.0135	0.0308	0.0549	0.0743	48.17	40.34	35.14	31.89

Tab. 2.1 - Ampiezza del quanto e SNR per quantizzazione ottima.

Analizzando tali valori, un primo risultato che si può ricavare è che il livello di saturazione ottimo (ottenibile moltiplicando l'ampiezza del quanto per il numero di livelli del quantizzatore) può essere fissato approssimativamente ad un valore pari a 4 σ_X. Per quanto riguarda il rapporto segnale/rumore di tale quantizzazione ottima, invece, si nota che i valori che si ottengono nei casi reali sono sensibilmente peggiori dei 6R dBm0 ottenuti nell'ipotesi di distribuzione uniforme con rumore di sovraccarico trascurabile. Di conseguenza a parità di distorsione, è necessario impiegare un quantizzatore con un maggior numero di bit per campione.

2.2.4 Caratteristiche del rumore di quantizzazione

Passando alle caratteristiche del rumore di quantizzazione, si vede come la sua funzione densità di probabilità possa essere considerata costante nell'intervallo $\pm \Delta/2$ (fig. 2.12). Infatti, indicando con $P(x_i)$ la probabilità di appartenenza al quanto i-esimo X_i ed essendo gli eventi di appartenenza a differenti quanti mutuamente esclusivi, la distribuzione di probabilità complessiva dell'errore p(e) è data dalla somma delle "n" distribuzioni all'interno dei singoli quanti $p(e_i)$ pesate tramite le $P(x_i)$

$$p(e) = \sum_{i=1}^{n} P(x_i) \cdot p(e_i) \tag{2.53}$$

La distribuzione di probabilità all'interno dei singoli quanti p(ei), d'altra parte, è legata a quella del segnale p(x). Infatti essa è pari alla distribuzione di probabilità p(x) condizionata dall'appartenenza a X_i, per cui vale

$$p(e_i) = \begin{cases} \dfrac{p(x)}{P(x_i)}; & x \in X_i \\ 0; & \text{altrimenti} \end{cases} \tag{2.54}$$

La p(e), quindi, si ottiene dalla somma di sezioni della p(x) (una per ciascun Xi), traslate centrandole nell'origine. Anche se le singole $p(e_i)$ risultano essere non costanti, la loro somma tende a distribuirsi uniformemente già per un numero di bit di quantizzazione $R \geq 2$, come mostrato in figura 2.12 per il caso di un segnale con distribuzione gaussiana e quantizzazione uniforme ottima.

Passando allo spettro dell'errore di quantizzazione, esso può essere stimato qualitativamente analizzando l'andamento del segnale nel tempo. Se il numero di livelli di quantizzazione risulta essere elevato ed il segnale non eccessivamente correlato, l'errore di quantizzazione è caratterizzato da rapidi cambiamenti di segno per il continuo attraversamento da parte del segnale dei differenti livelli di quantizzazione. Ciò porta ad una densità spettrale di potenza del segnale d'errore estremamente estesa approssimabile a quella di rumore bianco (fig. 2.13). Con tale ipotesi di distribuzione uniforme, noto che la potenza dell'errore di quantizzazione nella banda $\pm f_s/2$ è pari a $\Delta^2/12$, la sua densità spettrale di potenza Se(f) è pari a

$$S_e(f) = \frac{\Delta^2}{12}\frac{1}{f_s} \tag{2.55}$$

L'ipotesi fatta di quantizzazione con un numero elevato di livelli, però, è essenziale. Infatti, un'analisi della reale distribuzione spettrale di potenza del segnale d'errore mostra che essa non si mantiene costante, ma, come mostrato in figura 2.14, decresce all'aumentare della frequenza in maniera tanto più sensibile quanto meno elevato è il numero di bit del quantizzatore. L'approssimazione di spettro piatto, però, risulta sufficiente per il seguito della trattazione.

L'analisi precedente fatta sul segnale d'errore era relativa a segnali continui. Passando all'analisi dei segnali campionati, è conveniente ricavare lo spettro del rumore di quantizzazione campionato $e_c(t)$ come trasformata della

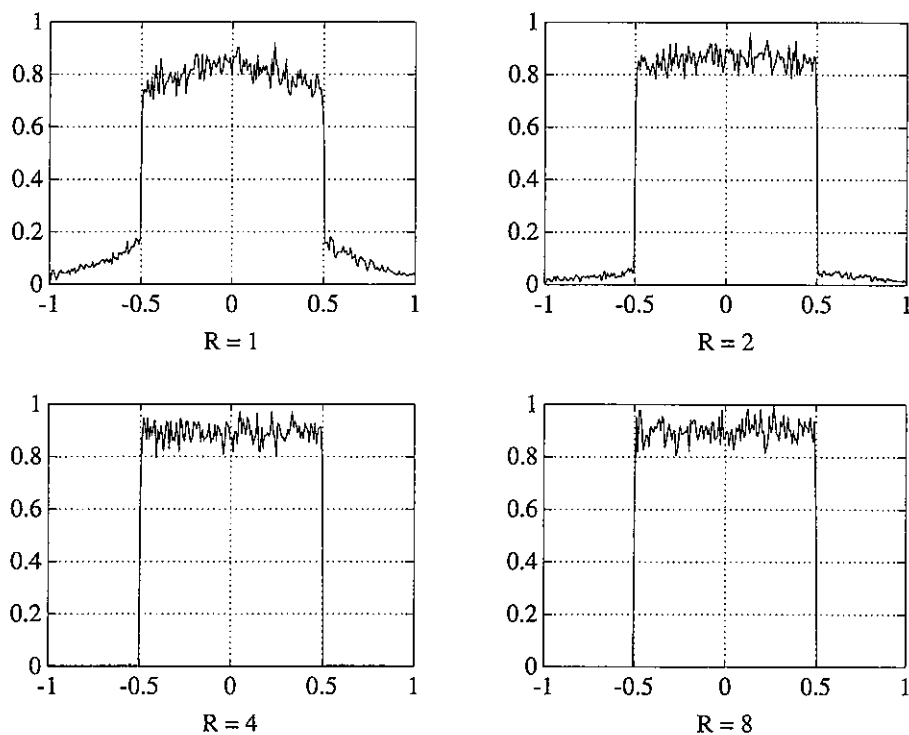



Fig. 2.12 - Distribuzione errore di quantizzazione.

sua funzione di autocorrelazione. Calcolando la funzione di autocorrelazione come media d'insieme, essa è esprimibile come

$$R_{ee}(\tau) \;=\; \int_{-\infty}^{\infty} e_c(t)\, e_c(t-\tau)\;\; p[\, e_c(t) = v,\, e_c(t-\tau) = v \,]\;\; dv \tag{2.56}$$

dove $p[\ e_c(t) = v,\ e_c(t\tau) = v\ ]$ rappresenta la probabilità condizionata che entrambi gli eventi assumano la generica ampiezza v. Gli estremi di integrazione si estendono per tutto l'intervallo di esistenza della funzione densità di probabilità del segnale d'errore. Per il calcolo dell'integrale è conveniente considerare i campioni dell'errore come ottenuti da rect(t) di larghezza infinitesima ε ed ampiezza e_g/ε, facendo poi tendere ε a zero

$$e_c(t_0) \;=\; \lim_{\varepsilon > 0} \frac{e_g(t_0)}{\varepsilon}\; \mathrm{rect}_{\varepsilon}(t_0) \tag{2.57}$$

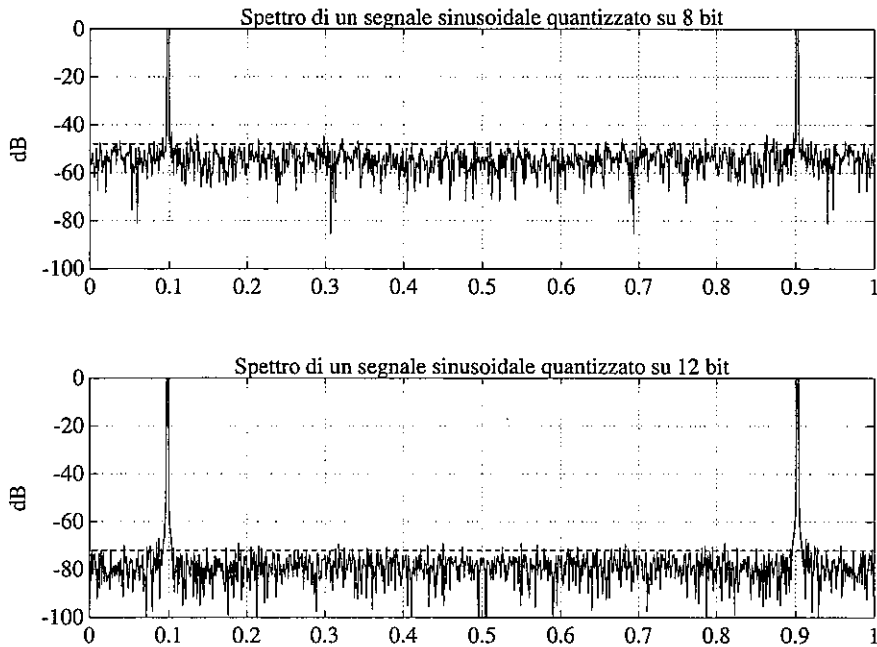



Fig. 2.13 - Spettri di segnali sinusoidali quantizzati.

Con tale ipotesi e considerando un periodo di campionamento T, la $p[\, e_c(t) = v,\, e_c(t-\tau) = v \,]$ può essere scomposta come prodotto della distribuzione di probabilità $p(e)$ della variabile $e_g(t)$ (identica per i due eventi) per la probabilità $p(v \mid t)$ che il segnale sia di ampiezza v nel generico istante t

$$R_{ee}(\tau) = \lim_{\varepsilon \to 0} \left[\int_{-\infty}^{\infty} e_g(t)\, e_g(t-\tau)\; p(e)\; p(v \mid \tau)\; dv \right] \tag{2.58}$$

Per il calcolo della $p(v \mid \tau)$, considerando la t come variabile aleatoria con distribuzione uniforme all'interno dell'intervallo di campionamento pari a 1/T, la probabilità che essa cada all'interno all'intervallo ε nel quale il segnale non è nullo è pari a ε/T. Inoltre, essendo il primo evento nullo per $t > \varepsilon$ ed il secondo nullo per $t > (\varepsilon - \tau)$, il calcolo della $p(v \mid \tau)$ va limitato nell'intervallo $[0, \varepsilon - \tau]$, ottenendo

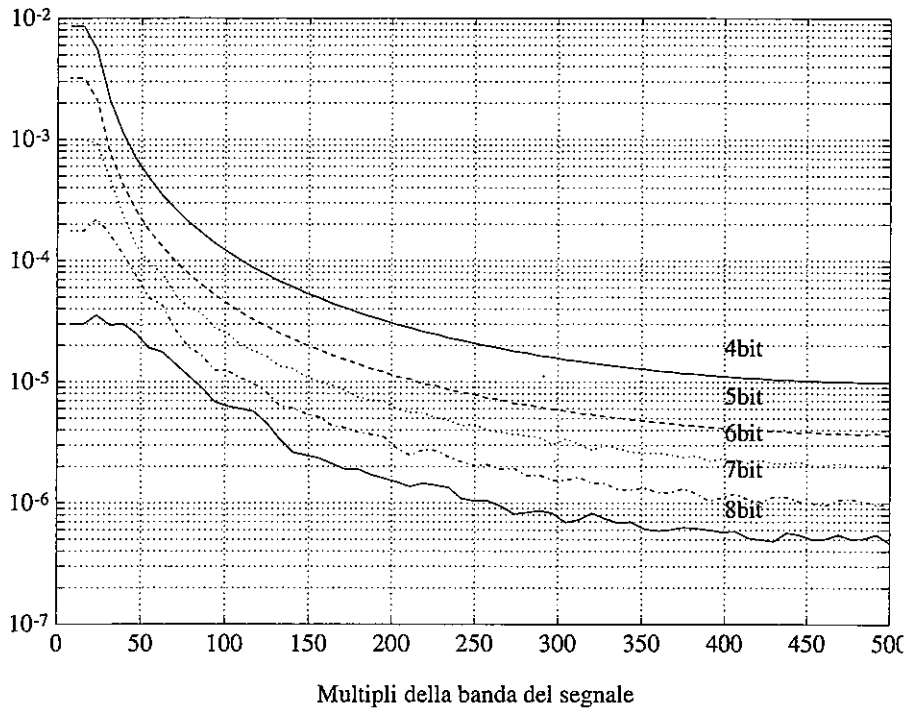



Fig. 2.14 - Densità spettrale di potenza di un segnale quantizzato.

$$p(v \mid \tau) = \int_0^{\varepsilon-\tau} \frac{\varepsilon}{T}\, dt = \frac{\varepsilon(\varepsilon-\tau)}{T} \tag{2.59}$$

$$R_{ee}(\tau) = \lim_{\varepsilon \to 0} \left[\int_{-\infty}^{\infty} e_g(t)\, e_g(t-\tau)\, \frac{\varepsilon(\varepsilon-\tau)}{T}\, dv \right] \tag{2.60}$$

Sempre nell'ipotesi di numero di livelli di quantizzazione elevato e segnale non fortemente correlato (o periodico), il campionamento dell'errore di quantizzazione alla frequenza di Nyquist porta a campioni tra loro scorrelati ed indipendenti dai campioni del segnale. L'ipotesi di indipendenza tra campioni dell'errore comporta che la funzione di autocorrelazione sia costituita da una d(t) centrata nell'origine. Infatti, solo facendo coincidere le due repliche del segnale utilizzate nel calcolo dell'autocorrelazione, i prodotti tra i vari campioni e, di conseguenza, la loro somma risultano sempre positivi; in tutti

gli altri casi, in media, il loro contributo tenderà ad annullarsi. Di conseguenza, riscrivendo l'espressione precedente per t = 0

$$R_{ee}(\tau) = \delta(\tau) \lim_{\varepsilon \to 0} \left[\int_{-\infty}^{\infty} e_c(t)^2 \, p(e) \; p(v \mid 0) \; dv \right]$$

$$= \delta(\tau) \lim_{\varepsilon > 0} \left\{ \int_{-\infty}^{\infty} \left[\frac{e_g(t)}{\varepsilon} \right]^2 p(e) \; \frac{\varepsilon^2}{T} \; dv \right\} = \frac{\delta(\tau)}{T} \int_{-\infty}^{\infty} [\, e_g(t) \,]^2 \, p(e) \; dv \tag{2.61}$$

Riconoscendo nell'argomento dell'integrale il valore quadratico medio e_g^2 della componente granulare del segnale d'errore, si ricava infine

$$R_{ee}(\tau) = \frac{e_g^2}{T} \, \delta(\tau) \tag{2.62}$$

Per il livello di quantizzazione i-esimo, utilizzando le relazioni già ricavate per e_g^2, ciò si traduce in

$$R_{xx}(e_i) = \frac{e_g^2}{T} \, \delta(\tau) = \frac{\delta(\tau)}{T} \, E\left\{ (x - x_i)^2 \right\} = \frac{\delta(\tau)}{T} \, \frac{\Delta^2}{12} \, P(x_i) \tag{2.63}$$

Integrando assumendo trascurabile la probabilità che il segnale cada al di fuori degli estremi di saturazione, si ottiene che la funzione di autocorrelazione complessiva

$$R_{ee}(\tau) = \frac{\delta(\tau)}{T} \, \frac{\Delta^2}{12} \tag{2.64}$$

La densità spettrale di potenza si ottiene trasformando tale funzione di autocorrelazione del processo a tempo discreto come

$$S(f) = \int_{-\infty}^{\infty} R_{ee}(\tau) \, e^{-j\frac{2\pi}{T}\tau} \, d\tau = \frac{1}{T} \, \frac{\Delta^2}{12} \tag{2.65}$$

dalla quale si osserva che lo spettro del segnale d'errore è uniforme (fig. 2.13). Integrando all'interno dell'intervallo ± 1/2T si ottiene che la potenza del segnale d'errore campionato nella banda di interesse, che vale

$$e_{\varepsilon}^{2} = \frac{\Delta^{2}}{12\,T^{2}} \tag{2.66}$$

Noto che la rappresentazione in frequenza di un segnale campionato Xc(f) è anch'essa scalata di un fattore pari a 1/T, il rapporto segnale/rumore ottenuto considerando segnali campionati

$$SNR = \frac{x_{\varepsilon}^{2}}{e_{\varepsilon}^{2}} = \frac{12}{\Delta^{2}}\, x_{\varepsilon ff}^{2} \tag{2.67}$$

coincide con quello ottenuto nel caso di segnali continui.

È opportuno, infine, analizzare con più dettaglio l'effetto della quantizzazione per ampiezze del segnale prossime allo zero. Si distinguono due tipi di caratteristiche di conversione (fig. 2.15). La prima è tale che i segnali con livello inferiore alla metà del passo di quantizzazione δ/2 vengono codificati con lo zero (codifica silenziata o "mid-tread quantizer").

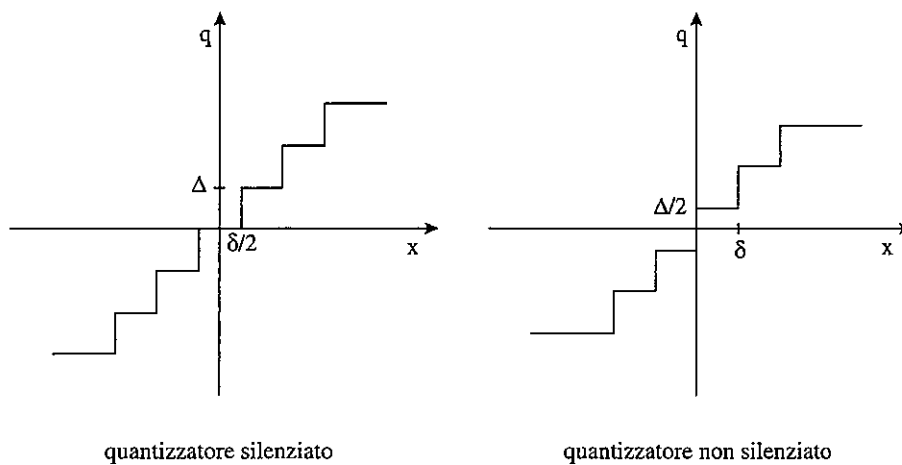



Fig. 2.15 - Quantizzatore silenziato e non silenziato.

Nella seconda lo zero viene scelto come livello di decisione (codifica non silenziata o "mid-riser quantizer"). La differenza tra i due quantizzatori risiede nel fatto che, mentre nel primo caso il rumore presente in assenza di segnale viene annullato (purché di ampiezza inferiore a δ/2), nel secondo caso

esso viene amplificato almeno a Δ/2. Entrambe le soluzioni hanno degli svantaggi. Con il codice silenziato si ha una riduzione della dinamica ai bassi livelli, con peggioramento della qualità del segnale. Con il codice non silenziato, invece, viene generato un rumore di fondo in assenza di segnale (idle channel noise) con potenza pari a (Δ/2)2. In realtà, dato che il corretto posizionamento del livello zero del segnale è difficilmente ottenibile (a causa di disturbi sul segnale o imperfezioni delle apparecchiature), esso può subire spostamenti dell'ordine della metà del passo di quantizzazione. La scelta di una caratteristica silenziata, quindi, viene vanificata ed è preferibile adottare una caratteristica non silenziata, che sfrutta tutti gli 'n' livelli permessi dalla dimensione del codice.

2.2.5 Convertitori A/D e D/A

La trattazione precedentemente svolta sulla rappresentazione numerica dei segnali analizzava separatamente gli effetti delle due trasformazioni di campionamento e quantizzazione. Nella pratica, le conversioni da analogico a digitale e da digitale ad analogico vengono, invece, eseguite in un'unica fase da componenti elettronici monolitici, detti convertitori [Tau77].

Analizzando la struttura dei convertitori A/D (fig. 2.16), comunque, si ritrovano due unità funzionali, che sono l'elemento di campionamento e tenuta (sample and hold: s&h) ed il quantizzatore, ricollegabili a tali trasformazioni. La funzione del s&h è quello di mantenere fissa la tensione ai capi del quantizzatore durante la conversione. Idealmente, esso potrebbe essere costituito dalla cascata di un interruttore e di una capacità. Nell'istante di campionamento l'interruttore si chiude per un tempo infinitesimo portando la capacità ad assumere ai propri capi una tensione pari a quella dell'ingresso. Aprendosi l'interruttore, la tensione ai capi della capacità viene mantenuta. In pratica, il comportamento di tali componenti non è ideale: l'interruttore è generalmente costituito da circuiti a transistor che rimangono chiusi per un intervallo finito di tempo, hanno una transizione tra apertura e chiusura graduale, il clock campionamento è soggetto a jitter, il trasferimento tra ingresso ed elemento di tenuta non è completo, l'elemento di tenuta è costituito da una circuiteria analogica soggetta a perdite di carica, ecc. Tutte queste cause fanno si che il funzionamento del s&h si discosti da quello teorico; l'errore

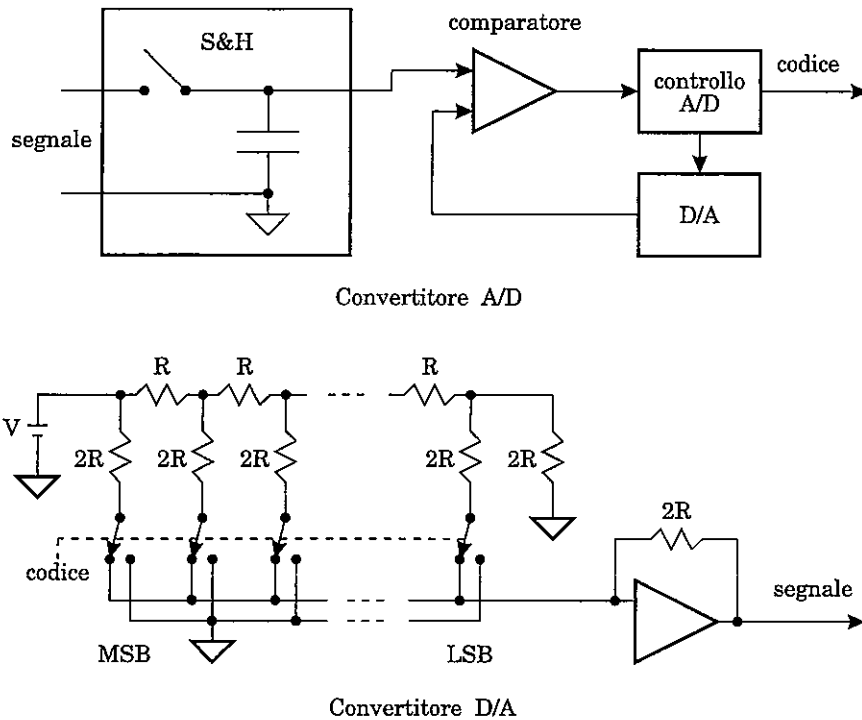



Fig. 2.16 - Struttura di convertitori A/D e D/A.

globale che ne deriva, però, viene mantenuto da parte del costruttore al di sotto dell'errore legato al numero di bit del quantizzatore.

Il quantizzatore può eseguire la conversione secondo diversi algoritmi, dei quali uno dei più diffusi è quello per approssimazioni successive. Con tale algoritmo i bit del codice vengono fissati in sequenza a partire dal più significativo (Most Significant Bit: MSB) procedendo verso il meno significativo (Least Significant Bit: LSB). L'algoritmo si basa sulla ricerca binaria del quanto di appartenenza del campione tra gli "n" presenti nel quantizzatore. Il MSB indica il segno del campione. Il MSB viene ottenuto confrontando il campione con un livello di prova di ampiezza nulla. Considerando, per semplicità, una codifica binary-offset (codice zero per il massimo negativo e con tutti i bit ad uno per il massimo positivo), il MSB viene posto ad uno o a zero a secondo che il campione risulti maggiore o minore del livello di prova. Il livello di decisione può essere generato tramite un

convertitore D/A opportunamente pilotato, mentre il comparatore può essere realizzato tramite circuiti utilizzanti amplificatori operazionali. Questa prima decisione individua nell'intero intervallo delle ampiezze comprese tra gli estremi di saturazione due semi intervalli, selezionando quello di appartenenza del campione. Il bit immediatamente meno significativo viene ottenuto confrontando il campione con un livello posto a metà del semi intervallo individuato con il MSB. Anche in questo caso, il bit viene posto ad uno o a zero a secondo che il campione risulti maggiore o minore del livello di prova. In tal modo si seleziona tra i quattro possibili intervalli di ampiezza pari ad un quarto della dinamica, quello al quale appartiene il campione. I rimanenti bit vengono determinati in maniera analoga, individuando di volta in volta intervalli di ampiezza via via dimezzata nel quale il campione è compreso. La codifica termina dopo R passi, dove R è il numero di bit.

Passando alla conversione D/A, il convertitori più comuni sono quelli a scala (fig. 2.16). Schematicamente, il segnale di uscita è generato da un operazionale configurato come sommatore. Gli R ingressi sono dati da correnti di ampiezza ordinatamente crescente, in modo tale che l'una risulti il doppio dell'altra. Ciò può essere ottenuto connettendo ad una sorgente di tensione un'opportuna rete resistiva [Tau77]. Nella somma, utilizzare o meno una corrente con un certo peso viene stabilito pilotando degli interruttori (reti a transistor) con i bit del codice.

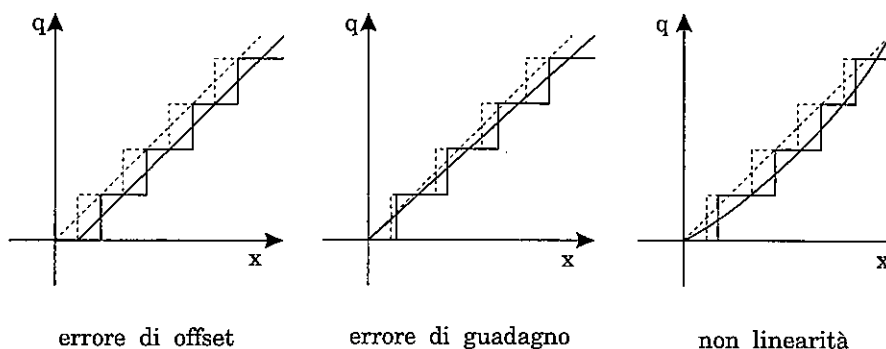



Fig. 2.17 - Distorsioni presenti nella conversione A/D.

Sia nella conversione A/D che nella conversione D/A sono presenti delle distorsioni (fig. 2.17), che sono la presenza di offset, errori di guadagno e non linearità. In presenza di offset, la caratteristica reale del convertitore risulta essere parallela a quella ideale, data dalla bisettrice del primo e terzo quadrante. In presenza di errori di guadagno, la caratteristica reale risulta essere una retta passante per l'origine, con inclinazione differente da quella della bisettrice. Nel caso di errori di non linearità, la caratteristica non è una retta ed i quanti, dunque, non risultano essere uniformi. Mentre i primi due tipi di errori sono legati alla circuiteria analogica del quantizzatore, il contributo principale all'ultimo tipo di distorsione è dato nella conversione A/D dall'errato posizionamento dei livelli di decisione, che è frutto dalla non proporzionalità della rete resistiva utilizzata nella conversione D/A. Specifico della conversione D/A è la presenza di impulsi spuri sul segnale analogico prodotto durante i transitori tra una conversione e l'altra. Ciò richiede l'adozione di reti di filtraggio che, unitamente ad altri problemi implementativi (es.: jitter, non linearità e rumore delle componenti analogiche, ecc.), portano a prestazioni dei convertitori A/D e D/A che possono scostarsi sensibilmente da quelle teoriche.

3

CODIFICA DI SORGENTE

3.1 RATE-DISTORTION FUNCTION

La conversione A/D trasforma una sorgente continua in un'equivalente sorgente discreta, a prezzo di una distorsione del segnale. Un'interpretazione di tale processo in termini di entropia associata alla sorgente continua H(x) ed all'uscita del quantizzatore H(y), vede il rumore di quantizzazione come equivocazione H(e) del mappaggio di intervalli del segnale continuo in ingresso al quantizzatore nei valori discreti dell'uscita

$$H(y) = H(x) - H(e) \tag{3.1}$$

Trasformata la sorgente da continua in discreta, è poi possibile applicare tecniche di codifica che riducano il numero di bit per campione R ad un valore prossimo a H(y). Si nota come il numero minimo di bit da utilizzare in una codifica di forma d'onda dipenda quindi dal livello di distorsione D: la R(D) è quindi detta rate-distortion function.

Per ottenere dei risultati quantitativi, è necessario esplicitare l'espressione dell'entropia. Nel caso di sorgente continua, ipotizzata senza memoria, essa è pari a

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) \, dx \tag{3.2}$$

Considerando, per semplicità, un segnale d'ingresso con distribuzione di probabilità gaussiana, l'entropia della sorgente è calcolabile come segue:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{x^2}{2\sigma_x^2}}$$

$$-\log_2 p(x) = -\frac{\ln p(x)}{\ln 2} = \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2} + \frac{x^2}{2\sigma_x^2}\right)$$

$$H(x) = \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2}\int_{-\infty}^{\infty} p(x)\,dx + \frac{1}{2\sigma_x^2}\int_{-\infty}^{\infty} p(x)\,x^2\,dx\right)$$

$$= \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2} + \frac{1}{2}\right) = \frac{1}{\ln 2}\left(\ln\sqrt{2\pi\sigma_x^2} + \ln e^{1/2}\right) = \frac{\ln\sqrt{2\pi e\sigma_x^2}}{\ln 2}$$

$$H(x) = \log_2\sqrt{2\pi e\sigma_x^2} \tag{3.3}$$

Tornando alla rate-distortion function, essa può essere descritta come

$$R(D) = \min[\,H(y)\,] = \min[\,H(x) - H(e)\,] = H(x) - \max[\,H(e)\,] \tag{3.4}$$

Analizzando il segnale di errore, dato che massimizzarne l'entropia equivale ad associare ad esso una distribuzione gaussiana ed indicando con $D = e_q^2$ la sua potenza, si ottiene

$$R(D) = \log_2\sqrt{2\,\pi\,e\,\sigma_x^2} - \log_2\sqrt{2\,\pi\,e\,D}$$

$$R(D) = \frac{1}{2}\log_2\frac{\sigma_x^2}{D} \tag{3.5}$$

Invertendo tale relazione si riottiene il legame già ricavato tra distorsione e numero di bit per campione

$$D(R) = 2^{-2R}\,\sigma_x^2$$

$$10\log_{10} D = -6R + 10\log_{10}\sigma_x^2 \tag{3.6}$$

Per segnali a distribuzione non gaussiana, la H(x) è inferiore a quella di segnali gaussiani, per cui l'espressione della R(D) ricavata ne rappresenta un limite superiore. D'altra parte, anche la distribuzione del segnale d'errore non necessariamente sarà gaussiana, per cui, il calcolo della rate-distortion function

come

$$R(D) = H(x) - \log_2 \sqrt{2 \pi e D} \tag{3.7}$$

fornisce, in realtà, il suo limite inferiore.

Passando al caso di sorgenti con memoria, l'entropia della sorgente è inferiore a quella di sorgenti senza memoria, con le relative ripercussioni sulla R(D). Tale aspetto, però, non viene ulteriormente approfondito in quanto, come mostrato nel seguito, nelle codifiche effettivamente implementate la correlazione tra i campioni del segnale viene solitamente eliminata tramite tecniche predittive, per cui l'ipotesi di sorgente discreta senza memoria torna ad essere valida.

3.2 CODIFICA ENTROPICA

Nella quantizzazione uniforme, la codifica dei campioni è a lunghezza fissa e avviene assegnando a ciascuno di essi la rappresentazione binaria dell'indice che li rappresenta tra L possibili valori generabili. Il numero di bit per campione R impiegato in tal modo è fisso e tale che

$$|\log_2 L| \leq R < |\log_2 L| + 1 \tag{3.8}$$

dove il segno di maggioranza è valido nel caso in cui L non sia una potenza di 2. In tal caso le dimensioni di R sono tali da permettere la codifica di un numero di simboli maggiori di quelli effettivamente utilizzati. Tale inefficienza della codifica a lunghezza fissa può essere ridotta considerando la codifica, sempre a lunghezza fissa, di blocchi di N simboli. Il numero di bit per blocco diventa pari a

$$N \cdot |\log_2 L| \leq R_N < N \cdot |\log_2 L| + 1 \tag{3.9}$$

e l'inefficienza di codifica, sempre inferiore ad un solo bit, viene in questo caso condivisa tra i simboli che appartengono al blocco, con un numero di bit per simbolo pari a

$$\begin{cases} R = \dfrac{R_N}{N} \\ |\log_2 L| \leq R_N < |\log_2 L| + \dfrac{1}{N} \end{cases} \tag{3.10}$$

Ad esempio, considerando una sorgente che emetta cinque simboli, la loro codifica è almeno di 3 bit. Se si considerano gruppi di due simboli, si ha un totale di 25 possibili gruppi, codificabili su 5 bit, con una media di 2.5 bit per simbolo.

Un quantizzatore, però, può essere considerato come una sorgente discreta che, in prima approssimazione, possiamo ipotizzare senza memoria. Fissata la distribuzione di probabilità del segnale in ingresso al quantizzatore e la sua caratteristica, è possibile calcolarne l'entropia come

$$H(x) = -\sum_{k=1}^{L} P_k \log_2 P_k \tag{3.11}$$

dove P_k rappresenta la probabilità del simbolo k-esimo (appartenente al quanto δ_k)

$$P_k = \int_{x \in \delta_k} p(x)\, dx \tag{3.12}$$

Se i simboli emessi non risultano equiprobabili, è possibile ridurre il numero medio di bit per simbolo utilizzando una codifica a lunghezza variabile di simboli o blocchi di essi. Questa codifica è tale da assegnare codici di lunghezza inferiore agli elementi emessi con maggiore frequenza dalla sorgente e viceversa. In tal modo si riduce il numero medio di bit per campione, definito come

$$\bar{R} = \sum_{k=1}^{L} P_k R_k \tag{3.13}$$

dove R_k rappresenta la lunghezza della codifica del simbolo k-esimo.

Per meglio comprendere il funzionamento di un codice a lunghezza variabile, è conveniente ricorrere ad una sua rappresentazione grafica (fig. 3.1). In una codifica a lunghezza fissa, le parole di codice possono essere viste come nodi terminali di un albero binario completo di ordine R. In una codifica a lunghezza variabile, invece, le parole di codice possono essere viste come nodi terminali di un albero sbilanciato, nel quale i percorsi più brevi sono assegnati ai simboli più probabili.

Affinché i simboli siano codificabili e decodificabili istantaneamente ed univocamente, è necessario che nessuna parola di codice sia la parte iniziale di

altre parole di codice (condizione del prefisso). Condizione necessaria e sufficiente affinché un codice a lunghezza variabile soddisfi la condizione del prefisso è che (disuguaglianza di Kraft)

$$\sum_{k=1}^{L} 2^{-R_k} \leq 1 \tag{3.14}$$

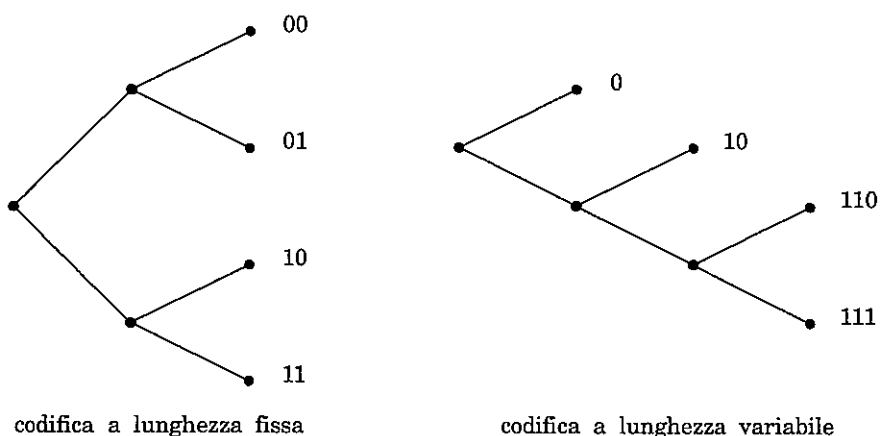



Fig. 3.1 - Rappresentazione grafica di codici a lunghezza fissa e variabile.

Per provare che la disuguaglianza di Kraft è una condizione sufficiente, si consideri il processo di costruzione dell'albero relativo ad una codifica a lunghezza variabile a partire da un albero binario bilanciato di ordine pari alla massima lunghezza di codice R_{MAX} (fig. 3.2). Il numero di nodi terminali di tale albero è pari a $2^{R_{MAX}}$. È poi possibile costruire un albero sbilanciato corrispondente ad una codifica a lunghezza variabile iniziando ad eliminare uno dei due possibili sotto-alberi relativi al livello 1 (dove il livello 0 è la radice dell'albero). In tal modo si è realizzato un nodo terminale, che corrisponde alla prima parola di codice, eliminandone contemporaneamente altri $2^{R_{MAX}-R_1}$ dall'albero completo. Continuando tale procedura, la frazione di nodi che vengono eliminati ad ogni passo (rispetto al totale dei nodi terminali) decresce progressivamente. D'altra parte, il codice generato assegnando a ciascuno dei nodi terminali che vengono via via creati una parola di codice soddisfa, per

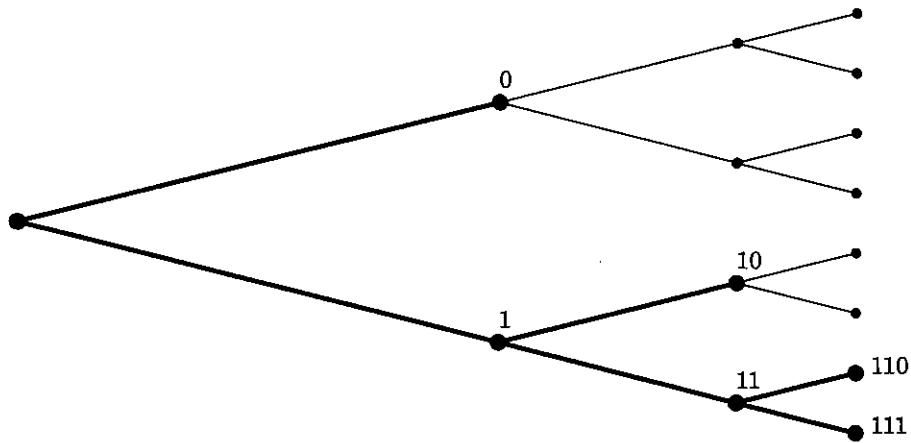



Fig. 3.2 - Costruzione di un albero relativo ad una codifica a lunghezza variabile a partire da un albero bilanciato.

costruzione, la condizione del prefisso. Notando che al livello j-esimo risulta

$$\sum_{k=1}^{j} \frac{2^{R_{MAX}-R_k}}{2^{R_{MAX}}} = \sum_{k=1}^{j} 2^{-R_k} < \sum_{k=1}^{L} 2^{R_k} \leq 1 \qquad (3.15)$$

risulta provato che la condizione è sufficiente. Per provare che la disuguaglianza di Kraft è una condizione necessaria, si osserva che i numeri di nodi eliminati al livello R_{MAX} è pari a

$$\sum_{k=1}^{L} 2^{R_{MAX}-R_k} \leq 2^{R_{MAX}} \qquad (3.16)$$

Dividendo tale relazione per il numero totale di nodi terminali si riottiene

$$\sum_{k=1}^{L} 2^{-R_k} \leq 1 \qquad (3.17)$$

Esempio: si consideri una sorgente discreta che emetta 4 simboli e se ne consideri una codifica a lunghezza variabile ottenuta secondo il procedimento descritto per la disuguaglianza di Kraft. La codifica risultante è riportata nella

seguente tabella:

	x_1	x_2	x_3	x_4
codifica	0	10	110	111
R_k	1	2	3	3

Tab. 3.1 - Codifica dei simboli emessi dalla sorgente.

Calcolando la sommatoria presente nella disuguaglianza di Kraft si ottiene

$$\sum_{k=1}^{4} 2^{-R_k} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{3} = 1 \qquad (3.18)$$

Tornando al problema di determinare il valor medio minimo di bit per campione, il teorema fondamentale della codifica per canali discreti in assenza di rumore (o primo teorema di Shannon) afferma che il minimo teorico di $\bar{R}$ coincide con il valore dell'entropia H(x) della sorgente. Considerando il caso di codifica di simboli isolati, è infatti, possibile dimostrare che

$$H(x) \leq \bar{R} < H(x) + 1 \qquad (3.19)$$

Nella dimostrazione di tale teorema, è utile considerare separatamente i limiti inferiori e superiori

$$\begin{cases} H(x) \leq \bar{R} \\ \bar{R} < H(x) + 1 \end{cases} \qquad (3.20)$$

Per il limite inferiore, che può essere riscritto come $H(x) - \bar{R} \leq 0$, si ha

$$H(x) - \bar{R} = \sum_{k=1}^{L} P_k \log_2 \frac{1}{P_k} - \sum_{k=1}^{L} P_k R_k = \sum_{k=1}^{L} P_k \log_2 \frac{2^{-R_k}}{P_k} \qquad (3.21)$$

Sfruttando la relazione

$$\ln x \leq x - 1 \qquad (3.22)$$

si ottiene

$$H(x) - \bar{R} \leq \log_2(e) \sum_{k=1}^{L} P_k \left(\frac{2^{-R_k}}{P_k} - 1 \right) = \log_2(e) \sum_{k=1}^{L} \left(2^{-R_k} - \frac{1}{P_k} \right) \quad (3.23)$$

Essendo, ovviamente, $P_k \leq 1$

$$\log_2(e) \sum_{k=1}^{L} \left(2^{-R_k} - \frac{1}{P_k} \right) \leq \log_2(e) \left(\sum_{k=1}^{L} 2^{-R_k} - 1 \right) \quad (3.24)$$

Riconoscendo, poi, nella sommatoria che appare nell'ultima equazione il primo membro della disuguaglianza di Kraft, rimane provato che

$$H(x) - \bar{R} \leq 0 \quad (3.25)$$

Per la dimostrazione del del limite superiore, è necessario esplicitare il tipo di codifica adottata. Nel caso di codifiche a lunghezza variabile univocamente ed istantaneamente decodificabili, la scelta degli interi R_k può essere fatta in modo tale che

$$2^{-R_k} \leq P_k < 2^{-R_k+1} \quad (3.26)$$

In tal modo si ha innanzitutto il vantaggio di legare la lunghezza della codifica alla probabilità del simbolo. Inoltre, tale codifica è valida in quanto, sommando il termine

$$2^{-R_k} \leq P_k \quad (3.27)$$

si riottiene la disuguaglianza di Kraft. D'altra parte, considerando il logaritmo del termine

$$P_k < 2^{-R_k+1} \quad (3.28)$$

si ottiene

$$\log_2 P_k < R_k + 1$$
$$R_k < 1 - \log P_k \quad (3.29)$$

Moltiplicando entrambi i membri di quest'ultima equazione per P_k e sommando si prova il limite superiore del primo teorema di Shannon

$$\bar{R} < H(x) + 1 \tag{3.30}$$

Passando al caso di codifica a lunghezza variabile di blocchi di simboli, l'inefficienza di codifica risulta ulteriormente ridotta in quanto, considerando blocchi di N simboli, risulta

$$N \cdot H(x) \geq \bar{R}_N > N \cdot H(x) + 1 \tag{3.31}$$

dove $\bar{R}_N$ è il numero medio di bit per blocco. Per il numero medio di bit per simbolo si ha

$$\begin{cases} \bar{R} = \dfrac{\bar{R}_N}{N} \\ H(x) \geq \bar{R} > H(x) + \varepsilon \end{cases} \tag{3.32}$$

dove ε è un numero positivo che può essere reso arbitrariamente piccolo. Ipotizzando di raggiungere un numero medio di bit per campione pari all'entropia della sorgente, il miglioramento che si otterrebbe rispetto all'uscita di un quantizzatore uniforme sarebbe pari a

$$\Delta R = \log_2 L - H(x) \tag{3.33}$$

Una codifica a lunghezza variabile notevolmente diffusa è la codifica di Huffman. Tale codifica viene eseguita ordinando i simboli emessi dalla sorgente discreta in una lista secondo le probabilità decrescenti. Innanzitutto si assegnano alla codifica dei due simboli con probabilità inferiore i bit "0" ed "1". Si considera poi l'unione dei due come un nuovo simbolo con probabilità pari alla somma delle loro probabilità, procedendo al riordino della lista. Si procede quindi all'iterazione di assegnazione di un bit alla codifica dei simboli meno probabili della lista ed al suo riordino fino a che tutti i simboli non abbiano avuto almeno due bit di codice.

Esempio: si consideri una sorgente che emetta otto simboli con la seguente probabilità:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
P	0.40	0.02	0.10	0.04	0.01	0.15	0.03	0.25

Tab. 3.2 - Probabilità dei simboli emessi dalla sorgente.

Se se ne calcola l'entropia si ottiene un valore pari a

$$H(x) = -\sum_{k=1}^{8} P_k \log_2 P_k$$

$$= -(0.40 \cdot \log_2 0.40 + 0.02 \cdot \log_2 0.02 + 0.10 \cdot \log_2 0.10 + 0.04 \cdot \log_2 0.04$$

$$+ 0.01 \cdot \log_2 0.01 + 0.15 \cdot \log_2 0.15 + 0.03 \cdot \log_2 0.03 + 0.25 \cdot \log_2 0.25)$$

$$H(x) = 2.29 \qquad (3.34)$$

La codifica di Huffman degli otto simboli si ottiene tramite le seguenti liste:

x_1 P=0.40	x_1 P=0.40	x_1 P=0.40	x_1 P=0.40	x_1 P=0.40	x_1 P=0.40	$x_{2,3,4,5,6,7,8}$ P=0.60	$x_{2,3,4,5,6,7,8}$ P=1.00
x_8 P=0.25	x_8 P=0.25	x_8 P=0.25	x_8 P=0.25	x_8 P=0.25	$x_{2,3,4,5,6,7}$ P=0.35	x_1 P=0.40	
x_6 P=0.15	x_6 P=0.15	x_6 P=0.15	x_6 P=0.15	$x_{2,3,4,5,7}$ P=0.20	x_8 P=0.25		
x_3 P=0.10	x_3 P=0.10	x_3 P=0.10	x_3 P=0.10	x_6 P=0.15			
x_4 P=0.04	x_4 P=0.04	$x_{2,5,7}$ P=0.06	$x_{2,4,5,7}$ P=0.10				
x_7 P=0.03	x_7 P=0.03	x_4 P=0.04					
x_2 P=0.02	$x_{2,5}$ P=0.03						
x_5 P=0.01							

Tab. 3.3 - Liste utilizzate nella codifica di Huffman.

Si ottiene la seguente codifica tramite codici a lunghezza variabile:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
codifica	1	0001010	0000	00011	0001011	001	000100	01
R_k	1	7	4	5	7	3	6	2

Tab. 3.4 - Codifica di Huffman.

Se si calcola la lunghezza media della codifica per simbolo

$$\bar{R} = \sum_{k=1}^{8} P_k R_k$$

$$= 0.40 \cdot 1 + 0.02 \cdot 7 + 0.10 \cdot 4 + 0.04 \cdot 5 + 0.01 \cdot 7 + 0.15 \cdot 3 + 0.03 \cdot 6 + 0.25 \cdot 2$$

$$\bar{R} = 2.34 \tag{3.35}$$

La codifica illustrata, oltre che applicata a singoli simboli emessi dalla sorgente discreta, può essere utilmente applicata anche a blocchi di essi.

Esempio: si consideri una sorgente discreta binaria caratterizzata dalla seguente distribuzione di probabilità di emissione dei due simboli

	x_1	x_2
P	0.80	0.20

Tab. 3.5 - Probabilità dei simboli emessi dalla sorgente.

L'entropia di tale sorgente è paria a

$$H(x) = -\sum_{k=1}^{2} P_k \log_2 P_k = -\left(0.80 \cdot \log_2 0.80 + 0.20 \cdot \log_2 0.20\right) = 0.72 \tag{3.36}$$

La codifica dei simboli isolati non può che avvenire che con un bit per simbolo. Se si considera, invece la codifica di Huffman di coppie di bit, si ottengono le seguenti liste

x(1,1) P=0.64	x(1,1) P=0.64	x(1,1) P=0.64	x(1,1),(1,2),(2,1),(2,2) P=1.00
x(1,2) P=0.16	x(2,1),(2,2) P=0.20	x(1,2),(2,1),(2,2) P=0.36	
x(2,1) P=0.16	x(1,2) P=0.16		
x(2,2) P=0.04			

Tab. 3.6 - Liste utilizzate nella codifica di Huffman.

che portano alla seguente codifica:

	x(1,1)	x(1,2)	x(2,1)	x(2,2)
codifica	0	11	100	101
R_k	1	2	3	3

Tab. 3.7 - Codifica di Huffman.

Il numero medio di bit per coppia di simboli è pari a

$$\overline{R}_N = \sum_{k=1}^{4} P_k R_k = 0.64 \cdot 1 + 0.16 \cdot 2 + 0.16 \cdot 3 + 0.04 \cdot 3 = 1.56 \tag{3.37}$$

e quindi un numero medio di bit per simbolo pari a $\overline{R} = 0.78$

3.3 CODIFICA DI SORGENTE ADATTATIVA

L'efficienza di una codifica di sorgente a lunghezza variabile è legata alla conoscenza a priori della probabilità dei simboli emessi. Nel caso in cui non si abbia tale informazione è necessario adottare algoritmi che estraggano in tempo reale le caratteristiche del flusso numerico emesso (compressione adattativa).

Per comprendere il funzionamento di una compressione adattativa è utile analizzare una ricodifica tramite Huffman di simboli codificati con parole di lunghezza fissa. In tal caso la compressione si ottiene riducendo la lunghezza dei codici più frequentemente emessi dalla sorgente ed aumentando la lun-

ghezza di quelli meno frequenti. Non disponendo della distribuzione della probabilità dei simboli, si può mantenere fissa la lunghezza delle parole di codice, ma associando ed esse blocchi di simboli di dimensione crescente in funzione della loro frequenza di emissione. In il numero medio di bit per simbolo, quindi, si riduce tanto più quanto maggiore è la possibilità di avere ripetizioni di lunghe sequenze di simboli.

Un algoritmo di compressione adattativa particolarmente diffuso è l'algoritmo di Ziv-Lemplel (LZ). Essendosi diffuso per applicazioni essenzialmente legate alla compressione di informazioni su memoria di massa, in tale algoritmo le sequenze a lunghezza variabile di simboli vengono dette "stringhe" e la dimensione dei codici, legata alla lunghezza delle parole nei sistemi di elaborazione, sono tipicamente a 12 o 16 bit.

Tale algoritmo si basa sulla realizzazione di una tabella di conversione, inizializzata per contenere, come stringhe iniziali, tutti i simboli generabili dalla sorgente (es.: tutti i possibili caratteri ASCII per un file di testo), presi singolarmente. Tale algoritmo di compressione si basa sulla proprietà per la quale ciascuna stringa presente nella tabella ha un prefisso "w" ottenuto eliminando il simbolo terminale "K" (estensione), anch'esso nella tabella.

Durante la compressione, vengono costruite stringhe di dimensioni crescenti ordinando sequenzialmente i simboli in ingresso (fig. 3.3). Contemporaneamente si controlla che la stringa w che si viene via via formando sia compresa in tabella, nel qual caso tale procedura continua. Nel momento in cui, ricevendo il simbolo K, la stringa wK non risulta presente nella tabella, si eseguono i seguenti passi:

- viene emesso il codice relativo alla stringa w, che rappresenta la più lunga delle stringhe già tabellate contenuta nella stringa da codificare;
- la stringa wK viene inserita nella tabella, assegnando ad essa un nuovo codice;
- K diventa il simbolo iniziale di una nuova stringa.

Analizzando le prestazioni di tale algoritmo, si nota come nella fase iniziale viene eseguita, in realtà, un'espansione del flusso (dato che singoli simboli vengono codificati in codici di dimensioni maggiori). Successivamente, però, ciascun codice emesso sarà via via relativo a stringhe di dimensioni crescenti, con un'efficienza che aumenta man mano che il processo di compressione procede.

Codifica

Tabella di codifica

stringhe	a	b	c	1b	2a	4c	3b	5b	8a	1a	10a	11a
codici	1	2	3	4	5	6	7	8	9	10	11	12

Decodifica

Fig. 3.3 - Algoritmo di Ziv-Lempel applicato ad una sequenza di tre simboli base.

Per la decodifica ciascun codice viene progressivamente scomposto nelle due componenti prefisso ed estensione. Tale scomposizione procede ricorsivamente fino a che il prefisso non rappresenti un simbolo isolato. Il simbolo finale di tale espansione viene utilizzato per aggiornate la tabella di decompressione, assegnando un nuovo codice alla stringa ottenuta dalla giustapposizione di tale simbolo alla stringa precedentemente ricevuta.

Trascurando il problema di avere in ricezione simboli ordinati inversamente rispetto alla trasmissione (risolvibile tramite una struttura LIFO), l'algoritmo di decompressione presentato fallisce qualora la stringa d'ingresso contenga la sequenza KwKwK, dove Kw appare già nella tabella di compressione. Il compressore, infatti, individuata la stringa Kw, invia il codice relativo ed aggiunge la stringa KwK alla sua tabella; successivamente, individuata la stringa KwK, utilizzerà il codice appena inserito. Tale codice risulta indefinito al decompressore, che risulta in attesa dell'estensione K alla stringa precedente. Per evitare tale inconveniente, ogni volta che in fase di decodifica ci si trova di fronte ad un codice non definito, si deve ipotizzare che tale codice sia un'estensione della stringa precedente, utilizzando come il primo simbolo da emettere l'estensione della stringa stessa.

Dal punto di vista implementativo, risulta conveniente memorizzare le stringhe d'ingresso come coppie di un codice (relativo al prefisso w della stringa corrente) ed un simbolo (relativo all'estensione k). Ciò permette l'accesso in tabella tramite codici di dimensioni fisse, potendo quindi adottare tecniche di hashing.

L'LZ ha molti vantaggi rispetto ad algoritmi che permettono livelli di compressione paragonabili. Innanzitutto non è legato all'eliminazione della ridondanza di un particolare tipo di sorgente; inoltre, la semplicità dell'implementazione ne permette l'impiego run-time. Come ulteriore punto a favore di tale algoritmo, si ha che esso non introduce degradazione del segnale, e quindi è utilizzabile per la compressione di dati numerici. Tra gli svantaggi si hanno:

- l'indeterminatezza delle dimensioni del flusso compresso (e quindi delle dimensioni della memoria di massa destinato a contenerlo o della banda necessario per trasmetterlo;
- la propagazione di errori di trasmissione;
- l'inefficienza della compressione per piccoli blocchi di dati;
- la perdita di efficienza per blocchi di grandi dimensioni contenenti dati eterogenei (dovuta all'omogeneizzazione della probabilità dei simboli).

Il fattore di compressione ottenibile con l'LZ è, ovviamente, funzione del livello di strutturazione del flusso in ingresso. Tale rapporto, che per testi alfanumerici o dati formattati supera il 100%, è normalmente prossimo al 65%.

3.4 QUANTIZZAZIONE VETTORIALE

Un risultato della codifica di sorgente è che è possibile ridurre l'inefficienza di codifica codificando blocchi di simboli piuttosto che simboli isolati. Questo approccio può essere seguito anche per quanto riguarda la quantizzazione. La quantizzazione uniforme precedentemente descritta prevede la codifica di campioni isolati del segnale, per cui viene anche indicata come quantizzazione scalare. La quantizzazione vettoriale, invece, prevede la codifica di segmenti di forma d'onda o, più in generale di insiemi di campioni o parametri, organizzati in un vettore.

Il processo di quantizzazione vettoriale può essere visto pertanto come l'associazione ad un vettore di ingresso X di un vettore riproduzione $Y(i)$ scelto da un insieme finito di elementi che prende il nome di vocabolario o codebook.

La codifica si traduce quindi nell'emissione di un indice i associato al vettore $Y(i)$ e rappresentativo del vettore di ingresso X. L'operazione di decodifica è duale e consiste semplicemente nell'indirizzamento di una tabella e la conseguente emissione del vettore indirizzato da i, il quale rappresenta la versione quantizzata del vettore di ingresso. Tali operazioni sono schematizzate in figura 3.4.

L'idea della Quantizzazione Vettoriale (QV) trae la sua origine dalla formulazione matematica di Shannon [Dha71] in cui un sistema di compressione di dati è modellato come un codice a blocco di sorgente, e dove i vari blocchi contigui e non sovrapposti sono trasformati in corrispondenti blocchi di simboli di canale. Tali codici di sorgente possono essere pensati come il risultato di una quantizzazione vettoriale (o multi-dimensionale), che comprende come caso particolare e sub-ottimo la quantizzazione scalare, in cui la lunghezza del vettore collassa ad uno.

Nonostante l'intrinseca superiorità della quantizzazione vettoriale rispetto a quella scalare in termini di efficienza di codifica, solo pochi lavori di ricerca sono apparsi in letteratura prima del 1979. Il motivo principale di tale apparente contraddizione risiedeva nella mancanza di uno strumento efficiente per la costruzione del vocabolario (o codebook).

Nel 1980 tre ricercatori americani: Linde, Buzo e Gray [Lin80], hanno proposto un algoritmo per la generazione del vocabolario il quale consente il progetto di quantizzatori vettoriali localmente ottimi con lunghezza del blocco, dimensione del vocabolario e misura di distorsione del tutto arbitrarie. Tale

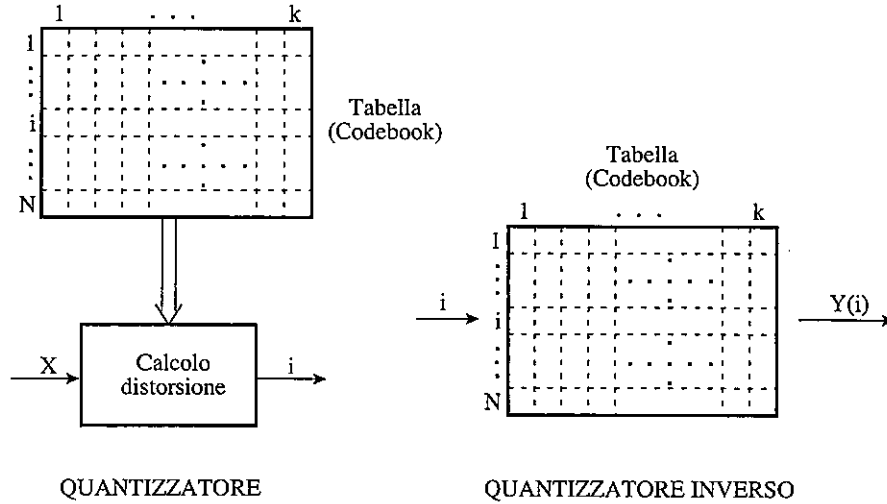



Fig. 3.4 - Schematizzazione delle procedure di quantizzazione vettoriale (codifica) e quantizzazione inversa (decodifica).

algoritmo ha preso successivamente il nome di algoritmo LBG, dalle iniziali dei tre ricercatori, ed ha determinato lo sviluppo di diversi lavori di ricerca nel settore, come pure l'applicazione della tecnica della quantizzazione vettoriale a vari schemi di codifica. In particolare la QV, tramite l'impiego di una particolare misura di distorsione, ha contribuito alla formulazione della tecnica di codifica CELP che sarà trattata diffusamente in un capitolo successivo.

Un quantizzatore vettoriale è un codificatore di sorgente caratterizzato da quattro elementi principali descritti nel seguito.

- Una sorgente discreta di vettori $\{\mathbf{X}\}$ appartenenti ad uno spazio reale a k-dimensioni $\mathbf{R}^k$ e cioè:

$$\mathbf{X} = \left[x_1, x_2, ..., x_k\right]^T \in \mathbf{R}^k \tag{3.38}$$

  In pratica tali vettori possono costituire segmenti successivi del segnale vocale campionato, raccolti in vettori di lunghezza k, ma possono anche essere rappresentativi di insiemi di coefficienti di predizione, o vettori di moduli di FFT ecc.

- Un insieme finito di vettori di riproduzione detto vocabolario (o codebook):

$$\mathbf{C} = \{\mathbf{Y}(i) \supset i = 1, 2, ..., N\} \quad (3.39)$$

in cui ogni vettore

$$\mathbf{Y}(i) = \left[y_1(i), y_2(i), ..., y_k(i)\right]^T \in \mathbf{R^k} \quad (3.40)$$

Ne consegue che il vocabolario di riproduzione sarà una matrice di $N \times k$ elementi.

- Una misura di distorsione $d(X, Y(i))$, anche detta misura di distanza, tra il generico vettore della sorgente X e la generica parola del vocabolario (codeword) $Y(i)$. Una misura di distorsione tra le più utilizzate è l'errore quadratico medio (mean squared error):

$$\text{m.s.e.} = \frac{1}{k} \sum_{j=1}^{k} \left(\mathbf{Y}_j(i) - \mathbf{X}_j\right)^2 \quad (3.41)$$

- Una regola di quantizzazione q(.) che mappa lo spazio dei reali a k dimensioni nel vocabolario

$$q : \mathbf{R^k} \rightarrow \mathbf{C} \quad (3.42)$$

La regola di quantizzazione universalmente utilizzata è quella della minima distorsione, e cioe:

$$q(\mathbf{X}) = \mathbf{Y}(i) \Leftrightarrow d(\mathbf{X}, \mathbf{Y}(i)) \leq d(\mathbf{X}, \mathbf{Y}(j)) \supset j \neq i \quad (3.43)$$

Come conseguenza diretta della introduzione della regola di quantizzazione, è possibile definire una partizione S dello spazio di ingresso come suddivisione dello spazio in sottoinsiemi s_i (cluster) ognuno dei quali contiene tutti i vettori di ingresso associati con la regola di quantizzazione q alla stessa parola del vocabolario $\mathbf{Y}(i)$. In formule:

$$S = \bigcup_i s_i \text{ con } s_i = \{\mathbf{X} \mid q(\mathbf{X}) = \mathbf{Y}(i)\} \supset i = 1, 2, ..., N \quad (3.44)$$

I parametri che definiscono le prestazioni di un quantizzatore vettoriale sono la *distorsione media* ed il rapporto di compressione che si suole esprimere in termini di *velocità di trasmissione*.

La *distorsione media* caratteristica di un quantizzatore vettoriale è un indicatore della fedeltà con cui i vettori dello spazio sorgente sono riprodotti ed è funzione della partizione S e del vocabolario di riproduzione C. In generale, la distorsione media può essere espressa come media di insieme dalla formula $D(S, C) = E[d(X, q(X))]$ dove E indica l'operatore di media di insieme e considerando che la partizione S è composta da N elementi disgiunti s_i si ottiene:

$$D(S,C) = \sum_{i=1}^{N} p(\mathbf{X} \in s_i)\ E\left[d(\mathbf{X}, \mathbf{Y}(i)) \mid \mathbf{X} \in s_i\right] \quad (3.45)$$

in cui $p(\mathbf{X} \in s_i)$ rappresenta la probabilità discreta che $\mathbf{X}$ appartenga al cluster s_i. Questa formulazione presuppone una conoscenza statistica della sorgente che spesso non è disponibile. Nei casi pratici si approssima la funzione densità di probabilità con d.d.p. note. Inoltre se il processo vettoriale costituito dalla successione dei vettori $\mathbf{X}(n)$ nel tempo è stazionario ed ergodico, la media di insieme può essere sostituita dalla media nel tempo a lungo termine, e quindi la relazione della distorsione media diventa:

$$D(S,C) = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{N} d(\mathbf{X}(n), q(\mathbf{X}(n))) \quad (3.46)$$

Quest'ultima formulazione è di gran lunga la più utilizzata ed in pratica vengono considerate per il progetto sequenze di addestramento (training) in cui M è sufficientemente grande e per le quali lo spazio sorgente è considerato rappresentativo della situazione operativa in cui il quantizzatore si troverà ad operare.

La *velocità di trasmissione* rappresenta il numero di bit necessari a trasferire l'informazione relativa all'indirizzo della parola di codice, quindi dipende dalle dimensioni del vocabolario, e vale:

$$R = \lceil \log_2 N \rceil \text{ bit/vettore} \quad (3.47)$$

È evidente quindi che per sfruttare al meglio la velocità di trasmissione, i vocabolari utilizzati avranno una dimensione potenza di due. Se si vuole

tenere conto anche della lunghezza del vettore si introduce la velocità espressa in bit/campione come:

$$r = \frac{1}{k} R = \frac{1}{k} \lceil \log_2 N \rceil \text{ bit/campione} \tag{3.48}$$

Quest'ultima relazione mette bene in evidenza la caratteristica dei quantizzatori vettoriali di consentire la codifica di segnali a velocità anche di frazioni di campioni al secondo. Questa caratteristica è molto utile nei sistemi di codifica a bassissima velocità o in quei sistemi in cui si vuole suddividere la velocità disponibile tra diversi segnali da codificare in modo ottimale.

Questi elementi sono sufficienti per implementare in modo semplice un algoritmo di quantizzazione vettoriale che nella sua realizzazione più diretta è un algoritmo di calcolo esaustivo della distorsione tra il vettore di ingresso da codificare e tutti quelli raccolti nel codebook. Il vettore del codebook a distorsione minima rappresenta la versione quantizzata del vettore di ingresso e l'indice che lo identifica rappresenta l'informazione da trasmettere al ricevitore.

Si noti infine che il vocabolario, pur essendo una matrice di valori reali, può essere memorizzato in virgola mobile, qualora la dinamica dei singoli elementi lo renda necessario, ma molto più spesso è rappresentato in memoria in aritmetica finita utilizzando un numero sufficiente di bit (tipicamente 16 bit). Questo ulteriore passo di quantizzazione può essere comunque completamente separato dal processo di quantizzazione vettoriale, nell'ipotesi in cui la distorsione introdotta sia trascurabile nei confronti di quella introdotta dal processo di QV.

3.4.1 Algoritmo LBG di generazione del vocabolario

Il problema principale nell'impiego della quantizzazione vettoriale consiste nel progetto del vocabolario, progetto che deve essere ottimo in relazione al criterio di minima distorsione. L'algoritmo LBG fornisce uno strumento efficace per la generazione di tale vocabolario.

Al fine di illustrare l'algoritmo, è necessario introdurre due criteri di ottimalità che fanno riferimento a due condizioni specifiche. L'algoritmo LBG utilizza tali criteri in un processo iterativo costruendo un vocabolario di volta in volta migliore dal punto di vista delle prestazioni.

Primo criterio di ottimalità

Si supponga di avere a disposizione un vocabolario di partenza $\mathbf{C}_0$ ma non la partizione $\mathbf{S}$ dello spazio sorgente. Tale partizione può essere costruita associando ogni vettore di ingresso $\mathbf{X}$ ad uno specifico vettore $\mathbf{Y}(i) \in \mathbf{C}_0$ scelto all'interno del vocabolario in modo da soddisfare il criterio di minima distorsione e cioè minimizzando $d(X, Y(i))$. Tale operazione consiste quindi nella quantizzazione di tutti i vettori dello spazio sorgente e nella loro suddivisione in opportuni sottoinsiemi (o cluster) relativi alla stessa parola del vocabolario. Se si indica tale partizione con $\mathbf{P}_{ott}$, ne consegue per come è stata costruita che

$$D(\mathbf{S}, \mathbf{C}_0) \geq D(\mathbf{P}_{ott}, \mathbf{C}_0) \tag{3.49}$$

e cioè la distorsione media associata alla partizione ottima è minore o al più uguale della distorsione media ottenibile con qualsiasi altra partizione.

Secondo criterio di ottimalità

In questo caso si supponga di avere a disposizione la partizione dello spazio sorgente $\mathbf{S}_0$ ma non il vocabolario $\mathbf{C}$. Per ogni sottoinsieme s_i della partizione esiste un vettore a minima distorsione $\bar{\mathbf{Y}}(s_i)$ tale da minimizzare la distorsione media nel singolo cluster s_i, e cioè tale che:

$$E\left\{ d(\mathbf{X}, \bar{\mathbf{Y}}(s_i)) \mid \mathbf{X} \in s_i \right\} \leq E\left\{ d(\mathbf{X}, \mathbf{Z}) \mid \mathbf{X} \in s_i \right\} \supset Z \tag{3.50}$$

Il vettore $\bar{\mathbf{Y}}(s_i)$ che soddisfa questa condizione prende il nome di *centroide* o centro di gravità generalizzato del cluster s_i. Supponendo ora di costruire un vocabolario $\mathbf{C}_{ott}$ come unione dei vari centroidi, otterremo il vocabolario ottimo $\mathbf{C}_{ott} = \left\{ \bar{\mathbf{Y}}(s_i);\ i = 1, 2, ..., N \right\}$ per il quale vale la relazione

$$D(\mathbf{S}_0, \mathbf{C}) \geq D(\mathbf{S}_0, \mathbf{C}_{ott}) \tag{3.51}$$

dove $\boldsymbol{C}$ rappresenta ogni possibile vocabolario.

Le due condizioni di ottimalità appena presentate sono gli elementi base utilizzati dall'algoritmo LBG che è un algoritmo iterativo strutturato sui seguenti quattro passi.

- Passo 1 - Dato un vocabolario iniziale $\mathbf{C_0}$ di dimensione N e con vettori di lunghezza k, una soglia di distorsione $\varepsilon \geq 0$ piccola a piacere, una funzione di distribuzione statistica della sorgente (o alternativamente una sequenza di addestramento sufficientemente lunga), si inizializza l'algoritmo con $m = 0$ (indice delle iterazioni) e $D_{-1} = \infty$ (distorsione media alla iterazione 1).
- Passo 2 - Dato il vocabolario corrente $\mathbf{C_m} = \{\mathbf{Y}(i);\ i = 1, 2, ..., N\}$, si applica il primo criterio di ottimalità calcolando la partizione ottima $\mathbf{P_m}(\mathbf{C_m}) = \{p_i;\ i = 1, 2, ..., N\}$. Si calcola quindi la distorsione media $D_m = D(\mathbf{P_m}(\mathbf{C_m}), \mathbf{C_m})$ associata alla partizione ottima $\mathbf{P_m}$ ed al vocabolario corrente $\mathbf{C_m}$.
- Passo 3 - Qualora la distorsione media sia diminuita percentualmente di una quantità inferiore ad ε nel passare dall'iterazione m-1 all'iterazione m e cioè in formule se

$$\frac{D_{m-1} - D_m}{D_m} \leq \varepsilon \tag{3.52}$$

  l'algoritmo termina con $\mathbf{C_m}$ vocabolario finale, altrimenti l'algoritmo continua con il passo 4.
- Passo 4 - Data la partizione ottima $\mathbf{P_m}$ calcolata al passo 2, si applica il secondo criterio di ottimalità calcolando il nuovo vocabolario $\mathbf{C_{m+1}} = \left\{ \hat{\mathbf{Y}}(s_i);\ i = 1, 2, ..., N \right\}$ come collezione dei centroidi di ogni cluster p_i. Si incrementa m e si ritorna al passo 2.

Il valore della distorsione media può essere calcolato come media di insieme, qualora si disponga della funzione di distribuzione statistica del segnale sorgente, o come media nel tempo qualora si disponga di una sequenza di addestramento sufficientemente lunga. Dalle due condizioni di ottimalità prima descritte, si evince che $D_m \leq D_{m-1}$ e quindi l'algoritmo converge ad un valore minimo seppur con un numero di iterazioni che può essere infinito. Linde, Buzo e Gray hanno dimostrato [Lin80] che il loro algoritmo tende a produrre un quantizzatore ottimo, se questo esiste, attraverso un metodo di successive approssimazioni, pertanto se l'algoritmo termina con un valore di $\varepsilon = 0$ in un numero finito di iterazioni, tale quantizzatore limite risulta determinato.

In realtà, a causa delle inevitabili approssimazioni introdotte (ad esempio considerando una sequenza di training finita), i quantizzatori vettoriali prodotti con questo metodo sono solo localmente ottimi. Tuttavia le prestazioni sono di gran lunga superiori a quelle ottenibili con quantizzatori scalari e consentono pertanto rapporti di compressione più alti.

A titolo di esempio, la figura 3.5a visualizza la procedura di generazione di un vocabolario tramite l'impiego dell'algoritmo LBG. L'esempio si riferisce ad un vocabolario di quattro parole con vettori di dimensione K=2 campioni che consentono una semplice rappresentazione grafica. La figura mostra l'insieme dei 4000 vettori che costituiscono la sequenza di addestramento, in questo caso relativa a campioni distribuiti uniformemente. I quattro punti in basso a destra rappresentano le parole iniziali del codebook, mentre gli asterischi indicano il codebook finale. Sono inoltre rappresentate le traiettorie degli spostamenti delle quattro parole all'aumentare delle iterazioni. L'andamento della distorsione nel passare dal codebook iniziale al codebook finale è riportato in figura 3.5b.

Le prestazioni, in termini di SNR e relative a QV con dimensioni e lunghezze dei vettori diverse, sono riportate in figura 3.6. Le prestazioni sono relative ad una sequenza di test diversa da quella utilizzata per l'addestramento e pertanto i risultati sono generalizzabili. In particolare in figura 3.6a sono riportate le prestazioni al variare della dimensione del vocabolario per lunghezze del vettore da 2 a 10. La lunghezza di addestramento era in questo caso costituita da circa 400.000 vettori di voce campionata a 8 kHz e relativa a 12 parlatori di tre lingue diverse. Si nota come le prestazioni crescano circa linearmente, in dB, con la dimensione. La figura 3.6b riporta i valori del grafico precedente parametrizzati rispetto alla velocità espressa in bit/campione. In questo caso appare evidente come vettori più lunghi consentano prestazioni migliori. In particolare si può osservare che circa le stesse prestazioni (10 dB) sono ottenibili con r=2 bit/campione e K=2 oppure con r=1 bit/campione e K=8 e cioè a velocità metà.

Resta infine da evidenziare che il calcolo dei centroidi dei singoli cluster è una operazione strettamente legata alla particolare misura di distorsione impiegata. Infatti esistono misure di distorsione per le quali il calcolo del centroide è particolarmente complesso o addirittura non definito. Nel caso specifico dell'errore quadratico medio, il centroide è semplicemente calcolato

come media aritmetica dei vettori appartenenti al cluster:

$$\hat{Y}(s_i) = | s_i |^{-1} \sum_{X \in s_i} X \tag{3.53}$$

3.4.2 Sequenza di addestramento

Come è stato già evidenziato, l'algoritmo LBG è solitamente usato impiegando una lunga sequenza di addestramento che consente di prescindere dalla conoscenza delle caratteristiche statistiche della sorgente. È evidente che tale sequenza di addestramento deve essere rappresentativa dello spazio sorgente, in caso contrario il quantizzatore vettoriale risulta subottimo e con prestazioni che sono molto variabili al variare del segnale in ingresso.

Una misura della significatività della sequenza di addestramento, consiste nella valutazione delle prestazioni di un quantizzatore vettoriale nei due casi distinti ottenuti considerando sequenze di test che fanno parte della sequenza di addestramento (inside) e sequenze che non sono state utilizzate per il progetto (outside). La sequenza sarà tanto più rappresentativa tanto minore è la differenza di prestazioni nei due casi inside e outside.

Il parametro che maggiormente influenza tale differenza di prestazioni è la lunghezza M della sequenza di addestramento, e cioè il numero di vettori considerati. Un esempio della differenza di prestazioni ottenibile, espresso in termini di SNR, è illustrato in figura 3.7, dove si vede come tale differenza diminuisca all'aumentare della lunghezza della sequenza.

In particolare nell'esempio specifico, relativo ad una VQ con k=8 campioni, si può notare che le prestazioni outside tendono a stabilizzarsi quando si sono usati più di 25000 vettori. Tenendo in conto che in questo caso la dimensione del QV è di 256 vettori, ne risulta che in media occorrono almeno 100 vettori per ogni singola codeword del vocabolario. Questa considerazione, relativa ad un esempio specifico, è tuttavia generalizzabile e solitamente si assume di avere a disposizione una sequenza di addestramento che contenga almeno 1000 vettori per codeword. Sempre facendo riferimento all'esempio di figura 3.7, considerando che la lunghezza del vettore è di 8 campioni, per un vocabolario di 4096 parole (R=12 bit/vettore o r=1.5 bit/campione) occorreranno circa 70 minuti di segnale vocale per progettare un QV con prestazioni

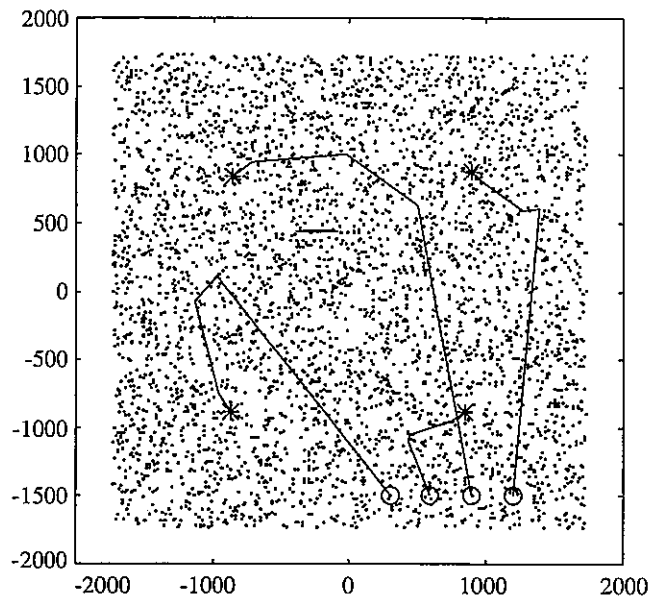



Fig. 3.5a - Procedura di generazione di vocabolario tramite LBG.

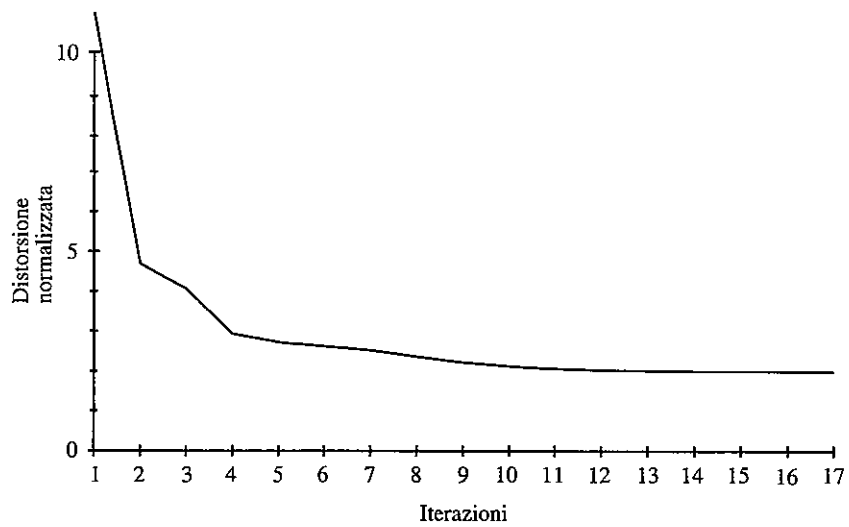



Fig. 3.5b - Esemplificazione della procedura di generazione di un vocabolario con k=2 e R=2.

adeguate. Vale la pena ancora sottolineare come le prestazioni valutate con sequenze che appartengono alla sequenza di addestramento siano decisamente fuorvianti in quanto, anche per sequenze di addestramento molto lunghe, risultano superiori a quelle effettive.

Un secondo elemento di importanza al fine delle prestazioni è costituito dalla necessità di considerare nella sequenza di addestramento segnali che siano rappresentativi delle condizioni in cui il QV opererà e quindi si dovrà tenere opportunamente in conto di diversi tipi di parlatori, diverse lingue, ecc. Nel caso poi di QV che devono operare su segnali preprocessati (come vettori relativi a set di coefficienti di filtri LPC, oppure insieme di moduli di una DFT), tutte le operazioni di preprocessamento che sono utilizzate nello schema di codifica, devono essere anche considerate per generare la sequenza di addestramento. Ne consegue che nel caso di schemi di codifica molto complessi, in cui il QV è inserito all'interno di un sistema in catena chiusa con delle possibili controreazioni, la sequenza di addestramento potrà dipendere dalla procedura di quantizzazione vettoriale stessa, e di conseguenza anche la generazione della sequenza di addestramento e la conseguente generazione del QV, dovranno essere condotte con procedure iterative, per approssimazioni successive.

3.4.3 Procedura di sdoppiamento

L'algoritmo LBG assume come condizione iniziale, oltre alla sequenza di addestramento, la conoscenza di un vocabolario di partenza $\mathbf{C}_0$. Tale condizione iniziale influisce in una certa misura sulla bontà del quantizzatore vettoriale prodotto dall'algoritmo stesso. Si può infatti intuire che se la funzione di costo associata alla generazione del vocabolario è una funzione complessa con molti minimi locali, il punto di partenza determina la zona nell'intorno della quale l'algoritmo troverà il minimo.

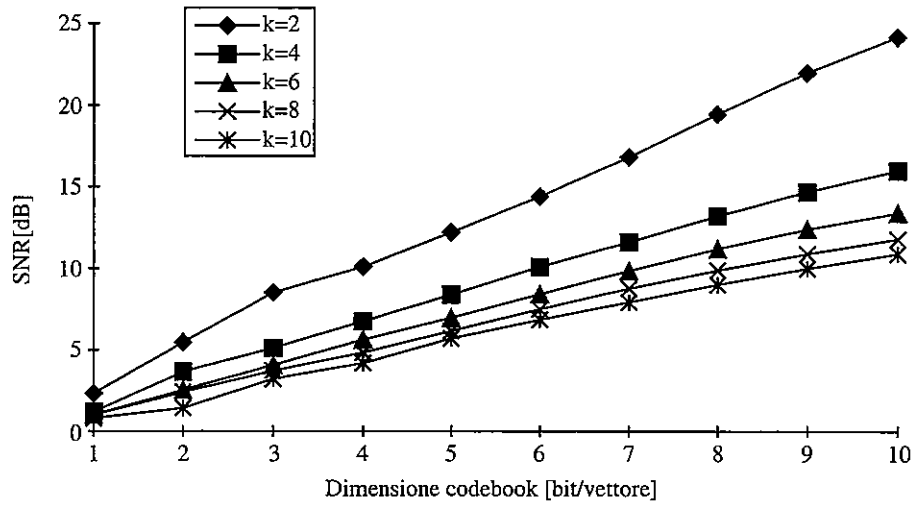
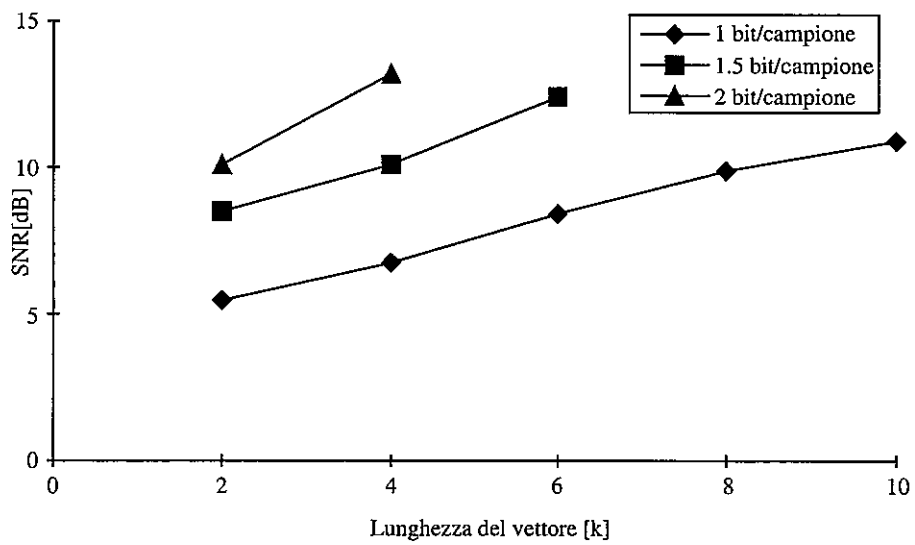
Inoltre una scelta errata del vocabolario di partenza può determinare malfunzionamenti dell'algoritmo LBG come l'occorrenza di celle vuote. Questo fenomeno si presenta in quei casi in cui nella determinazione della partizione ottima dello spazio sorgente (primo criterio di ottimalità), per certe codeword non esistono vettori dello spazio sorgente associabili tramite il criterio della minima distorsione. Questo fenomeno in realtà è considerato problematico anche in quei casi in cui il numero di vettori di ingresso associati ad una codeword sia particolarmente esiguo (meno del 5 % del valore medio

(a)

(b)

Fig. 3.6 - Andamento del rapporto segnale su rumore (SNR): (a) al crescere della dimensione del vocabolario N e per diverse lunghezze del vettore k, (b) al crescere della lunghezza del vettore e per diverse velocità r in bit/campione.

delle altre codeword). Si intuisce come un vocabolario in cui alcune parole siano rappresentative di una porzione molto ristretta dello spazio sorgente abbia prestazioni medie subottime. Per completezza va detto che questo fenomeno può anche essere indicativo di una sequenza di addestramento troppo corta, sicuramente non tale da consentire le stesse prestazioni inside ed outside.

La tecnica più utilizzata per ovviare a questo problema è quella dello sdoppiamento o *splitting procedure* che riduce la ricerca del vocabolario iniziale alla determinazione di una sola codeword di partenza. Questo corrisponde a considerare un vocabolario iniziale con una sola parola e tale parola può quindi coincidere con il centroide della intera sequenza di addestramento.

La procedura di generazione di un vocabolario di dimensione N avviene quindi tramite i seguenti passi:

- Passo 1 - Si calcola il centroide dell'intera sequenza di addestramento $\mathbf{Y_c}$ e si produce un vocabolario iniziale di $n = 2$ parole moltiplicando il centroide per due vettori di perturbazione η_1 e η_2:

$$\mathbf{C}_0^{\{n\}} = \{\mathbf{Y_c} \cdot \eta_1\} \cup \{\mathbf{Y_c} \cdot \eta_2\} \tag{3.54}$$

- Passo 2 - Si calcola il vocabolario ottimo $\mathbf{C}_m^{\{n\}}$ relativo alla dimensione n utilizzando l'algoritmo LBG prima descritto
- Passo 3 - Se la dimensione n è quella voluta, il vocabolario finale è prodotto, altrimenti si raddoppia la dimensione del vocabolario utilizzando la splitting procedure:

$$\mathbf{C}_0^{\{n \cdot 2\}} = \{\mathbf{C}_0^{\{n\}}(\eta_1)\} \cup \{\mathbf{C}_0^{\{n\}}(\eta_2)\} \tag{3.55}$$

  in cui i due semivocabolari sono ottenuti moltiplicando ogni singola codeword del vocabolario ottimo di dimensione n per i vettori di perturbazione η_1 e η_2 rispettivamente. Quindi si raddoppia il valore di n e si ritorna al passo 2.

La procedura appena descritta si presta molto bene al progetto di quantizzatori vettoriali con dimensioni potenza di 2, che rappresentano tuttavia la quasi totalità dei casi per gli evidenti vantaggi in termini di efficienza di trasmissione.

La scelta dei valori da utilizzare per i vettori di perturbazione η_1 e η_2 dipende dalla particolare sorgente, ma in generale non è molto critica e

solitamente questi hanno valori nell'intervallo 0.01 0.1. Inoltre al fine di garantire una diminuzione della distorsione media nel passare da una dimensione alla dimensione doppia, si pone uno dei due vettori di perturbazione pari al vettore unitario. Questo garantisce che il vocabolario a dimensione doppia contenga una replica del vocabolario ottimo calcolato al passo precedente, più una sua versione modificata.

Sebbene la procedura di sdoppiamento costituisca uno strumento efficce per risolvere il problema delle celle vuote (o sottopopolate), in alcuni casi particolari il problema può persistere ed in questi casi si suole inserire nell'algoritmo di progetto un ulteriore passo di controllo in cui viene valutato il numero di vettori presenti in ogni cluster. Qualora per qualche cella questo sia minore di una certa soglia percentuale, si introduce uno sdoppiamento di quelle celle che maggiormente contribuiscono alla distorsione media.

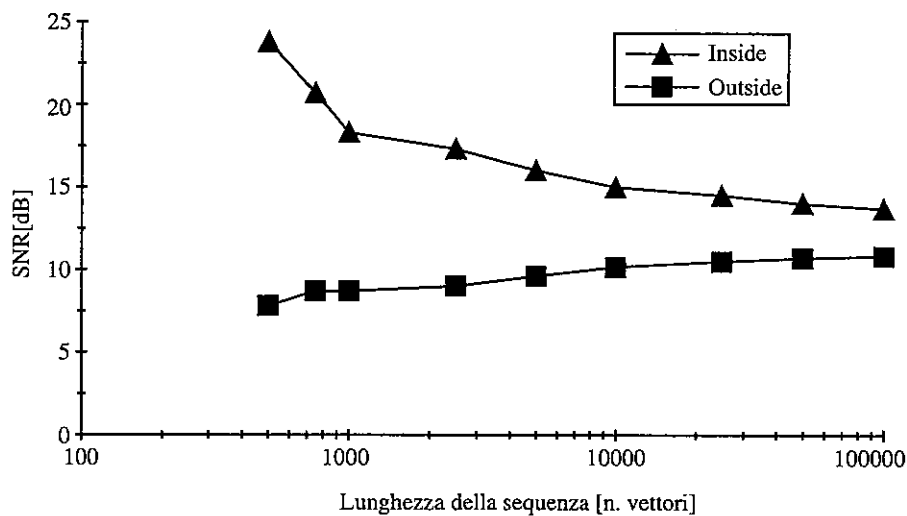



Fig. 3.7 - Andamento della distorsione media tra sequenze inside e outside alla sequenza di training, all'aumentare della lunghezza della sequenza di addestramento.

3.4.4 Struttura dei quantizzatori vettoriali

Nella presentazione della tecnica di quantizzazione vettoriale così pure come nella descrizione dell'algoritmo LBG, abbiamo fino ad ora ipotizzato che il processo di quantizzazione avvenga confrontando il vettore di ingresso con tutti i

possibili vettori disponibili nel vocabolario, al fine di determinare quello a distorsione minima. Tale tipo di ricerca esaustiva (full-search), seppur molto diffusa, è solo una delle possibili tecniche di ricerca e quindi di quantizzazione vettoriale.

Un'altra possibilità introdotta in letteratura è quella della ricerca ad albero (tree-search) che ha come caso particolare la ricerca binaria (binary-search).

Il motivo per cui si sono considerate tecniche alternative di ricerca, risiede nella complessità di calcolo associata con la tecnica di QV a ricerca esaustiva. Infatti un algoritmo di ricerca completa richiede per la quantizzazione di un vettore, il calcolo di N valori di distorsione dove N è il numero di parole del vocabolario. Nel caso anche semplice di misura di distorsione come il m.s.e., sono necessarie almeno k somme e k prodotti per ogni vettore e quindi per QV con lunghezze di vettori superiori a 20 e dimensioni superiori a 12 ÷ 14 bit la complessità può cominciare ad essere un problema.

La tecnica di ricerca ad albero si basa sulla costruzione di un vocabolario strutturato ad albero in cui quindi il vettore di ingresso è confrontato con un sottoinsieme di codeword e quella a distorsione minima individua la strada da percorrere all'interno dell'albero per giungere al vettore rappresentativo che è contenuto nell'ultimo livello dell'albero. Questa procedura è schematizzata in figura 3.8 per un esempio di quantizzatore vettoriale a tre livelli con dimensioni pari a 2 per il primo livello, 4 per il secondo e 2 per il terzo.

Nell'esempio in questione il vettore di ingresso è confrontato con le due parole al primo livello. Si supponga la parola a distorsione minore sia $\mathbf{Y}_{1,0,0}$, questa individua i quattro vettori al livello 2 tra cui calcolare la codeword più rappresentativa. Una volta selezionato il vettore $\mathbf{Y}_{1,2,0}$ al secondo livello, l'ultima ricerca verrà effettuata al terzo livello tra i due vettori $\mathbf{Y}_{1,2,1}$ e $\mathbf{Y}_{1,2,2}$ e nell'ipotesi il vettore a minima distanza sia $\mathbf{Y}_{1,2,2}$, questo costituirà la versione quantizzata del vettore di ingresso. La dimensione massima del vocabolario è in questo esempio di 16 parole (tutte quelle dell'ultimo livello), ma il numero di calcoli di distorsione è ridotto a otto, in contrapposizione ai sedici necessari con una procedura full-search.

Dall'esempio appare anche evidente che, a fronte della riduzione di complessità di calcolo, la struttura ad albero comporta un aumento della quantità di memoria necessaria per memorizzare il vocabolario che in questo caso passa da 16 celle a 26. Si noti tuttavia che tale aumento di capacità di memoria esiste solo al trasmettitore, in quanto al ricevitore è sufficiente memorizzare l'ultimo livello e quindi ancora solo 16 celle.

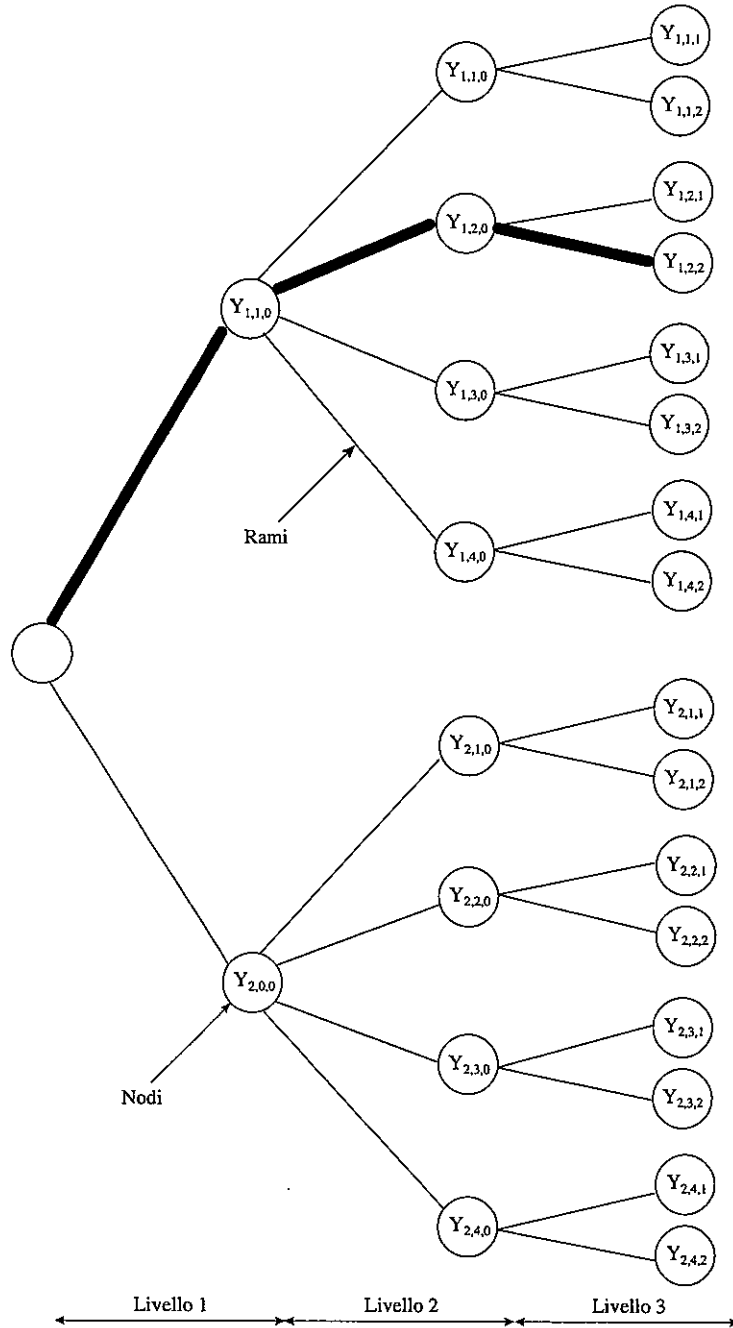



Fig. 3.8 - Struttura di un quantizzatore vettoriale ad albero a tre livelli.

La velocità complessiva, nel caso di struttura ad albero, sarà data dalla somma delle velocità parziali ad ogni singolo livello:

$$R = \sum_{n=1}^{L} R_n \tag{3.56}$$

In generale quindi si può concludere che i QV strutturati ad albero consentono una riduzione di complessità di calcolo a scapito di un aumento della quantità di memoria necessaria per memorizzare esplicitamente il vocabolario. Tale riduzione di complessità è tanto maggiore quanto minore è il numero di rami che si dipartono da ogni singolo nodo. Questa riduzione diventa quindi massima nel caso di struttura binaria dove da ogni nodo si dipartono due soli rami. Il confronto in termini di complessità e di memoria tra le tre strutture descritte è riportato in tabella 3.8 dove R rappresenta la velocità complessiva del QV in bit/vettore ed L il numero di livelli.

	Numero di livelli	Velocità al livello n	Numero di calcoli di distorsione	Numero di vettori da memorizzare
Ricerca esaustiva	1	R	2^R	2^R
Ricerca ad albero	L	R_n	$\sum_{n=1}^{L} 2^{R_n}$	$\sum_{i=1}^{L} \prod_{n=1}^{i} 2^{R_n}$
Ricerca binaria	L=R	1	$2 \cdot R$	$\sum_{i=1}^{R} 2^i$

Tab. 3.8 - Confronto in termini di complessità di calcolo e memoria necessaria tra le diverse strutture di quantizzatori vettoriali.

Il progetto di QV con struttura ad albero utilizza l'algoritmo base LBG con alcune semplici modifiche. Si comincia calcolando il QV, e quindi il vocabolario ottimo, per il primo livello. Quindi si divide la sequenza di addestramento nei vari cluster relativi ad ogni singolo nodo di tale livello (calcolo della partizione ottima). Per ogni nodo, ed utilizzando come sequenza di addestramento i vettori del cluster, si calcola il vocabolario ottimo di dimensione pari al numero di rami che si dipartono da quel livello. La procedura continua fino all'ultimo livello allo stesso modo.

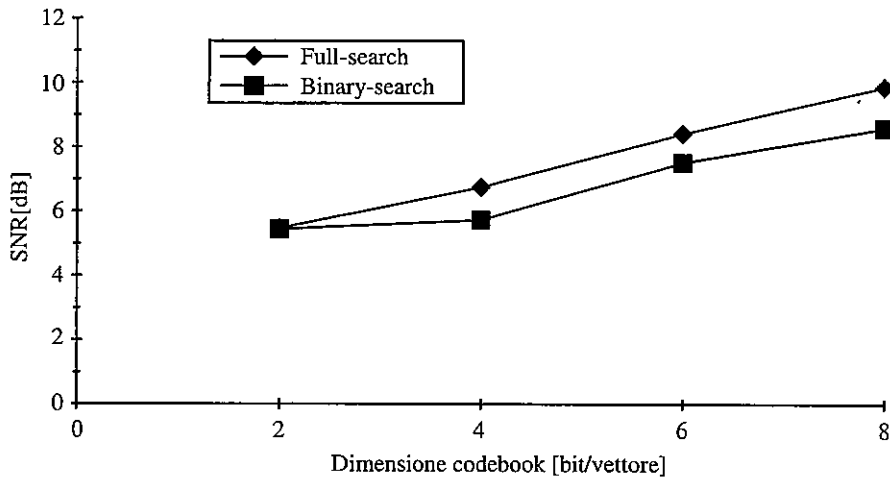



Fig. 3.9 - Confronto di prestazioni tra struttura full-search e binary-search. La velocità è r=1 bit/campione e quindi la lunghezza del vettore k coincide con la dimensione del code book R.

Relativamente alle prestazioni, tutte le strutture di ricerca semplificate comportano una, seppur contenuta, degradazione rispetto al caso di ricerca esaustiva. A titolo esemplificativo la figura 3.9 riporta le prestazioni ottenibili, a parità di velocità di trasmissione, utilizzando una struttura full-search ed una struttura tree-search. L'esempio si riferisce ad uno spazio sorgente costituito da vettori di segnale vocale campionato ad 8 kHz

Si è visto quindi che le strutture ad albero consentono una riduzione della complessità di calcolo a scapito di un aumento della capacità di memoria. Le strutture cosiddette multistadio consentono invece un risparmio anche della capacità di memoria.

Il principio su cui si basano queste strutture è quello di operare una seconda QV sul segnale errore di quantizzazione prodotto da un primo QV. Questa struttura è rappresentata in figura 3.10.

In generale i due QV avranno dimensioni N1 ed N2. Ipotizzando siano strutturati entrambi in full-search ed abbiano vettori della stessa lunghezza k, la quantità di memoria necessaria sarà di mem = $(N_1 + N_2)$ vettori reali e questo coincide anche con il numero di calcoli di distorsione necessari calc = $(N_1 + N_2)$. La velocità complessiva sarà data da $R = R_1 + R_2 = \log_2 N_1 + \log_2 N_2$.

Un QV a singolo stadio a pari velocità di trasmissione, anch'esso con struttura full-search, avrà velocità pari a $R = R_1 + R_2$ e quindi un numero di

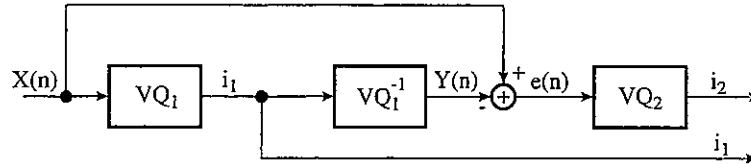



Fig. 3.10 - Struttura di un quantizzatore vettoriale multistrato a due stadi.

parole nel vocabolario pari a $N = 2^R = 2^{R_1} \cdot 2^{R_2} = N_1 \cdot N_2$. La memoria necessaria ed il numero di operazioni quindi in questo caso saranno pari a $N_1 \cdot N_2$. Ne consegue che il fattore di risparmio γ, nell'utilizzare una struttura multi- stadio, vale:

$$\gamma = \frac{N_1 \cdot N_2}{N_1 + N_2} \tag{3.57}$$

Questo risparmio può essere consistente e per QV di grande dimensione si possono considerare anche più stadi in cascata. Anche in questo caso a fronte di significativi vantaggi computazionali si ha una degradazione delle prestazioni. Un confronto di prestazioni è fornito in figura 3.11. Nel caso multiple stage sono stati considerati due stadi con uguale dimensione.

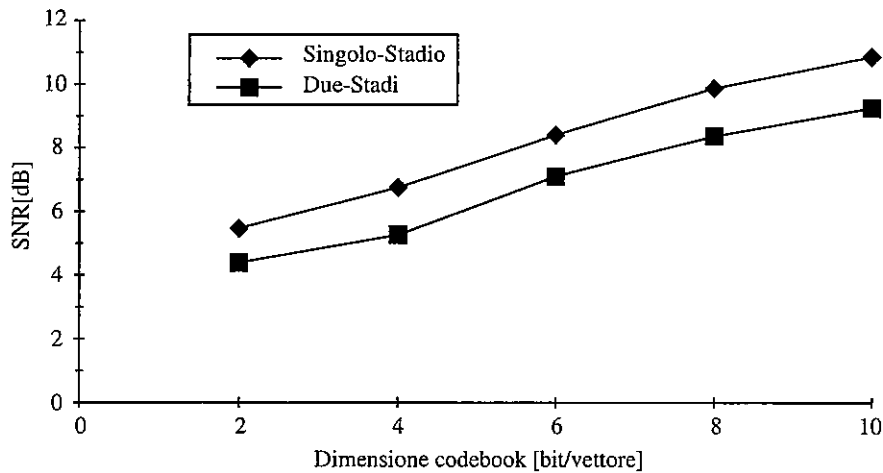



Fig. 3.11 - Confronto di prestazioni tra struttura a singolo stadio e struttura a due stadi. La velocità è r=1 bit/campione e quindi la lunghezza del vettore k coincide con la dimensione del code book R.

4

CODIFICA NUMERICA DI FORMA D'ONDA SENZA MEMORIA

4.1 COMPRESSIONE PER QUANTIZZAZIONE NON UNIFORME

Per una codifica numerica tramite quantizzazione uniforme, fissati la frequenza di campionamento fc (in funzione della banda del segnale), gli estremi di saturazione V (nota la distribuzione di probabilità delle ampiezze) ed il numero dei livelli (in funzione della dinamica desiderata) del quantizzatore, rimangono fissati il numero dei campioni al secondo generati dalla sorgente ed il numero di bit per campione R con cui essi vengono codificati. Di conseguenza è fissata la velocità fN del flusso numerico prodotto dalla sorgente stessa ($f_N = f_c \times R$). Ad esempio, nel caso del segnale telefonico, trasmissibile analogicamente con una banda di 4 kHz, usando una frequenza di campionamento di 8 kHz ed una quantizzazione su 12 bit, viene generato un flusso di 8 * 12 = 96 kbit/s: per la trasmissione di questo flusso può essere richiesta una banda superiore ai 40 kHz.

Considerando un segnale audio con banda di 15 kHz quantizzato su 16 bit ed campionato a 32 kHz (es.: DAB), flusso che viene generato è di 512 kbit/s. Nel caso di segnale audio con banda di 20 kHz quantizzato su 16 bit, se si utilizza un campionamento a 44.1 kHz (es.: CD) il flusso generato è di 705 kbit/s, mentre con un campionamento a 48 kHz (es.: DAT) si ha un flusso di 768 kbit/s. Non considerando l'incremento dovuto all'aggiunta di codici per la correzione degli errori di trasmissione, tale flusso si raddoppia nel caso di segnali stereofonici, per cui la velocità di riferimento per un segnale audio digitale a larga banda è di 1.5 Mbit/s.

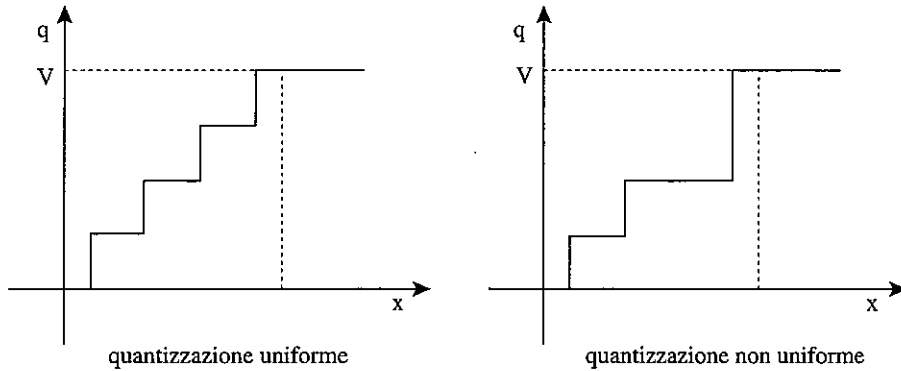



Fig. 4.1 - Riduzione dei livelli del quantizzatore con quantizzazione non uniforme.

È evidente che, trascurando altri aspetti (robustezza agli errori di trasmissione, semplificazione degli apparati, predisposizione all'elaborazione numerica, ecc.), sarebbe auspicabile adottare qualche forma di compressione. Volendo ridurre tale flusso, nelle codifiche di forma d'onda non è possibile ridurre il numero di campioni al secondo (riducendo la frequenza di campionamento tramite decimazione del segnale), per la riduzione della banda utile del segnale che ne conseguirebbe. È necessario, quindi, ridurre il numero di bit per campione riducendo i livelli del quantizzatore.

La riduzione dei livelli del quantizzatore può avvenire o riducendo gli estremi di saturazione del quantizzatore (come descritto a proposito della codifica di forma d'onda con memoria) o adottando, dove possibile, quanti di ampiezza maggiore (quantizzazione non uniforme) (fig. 4.1). Ciò può avvenire secondo due criteri: il primo (quantizzazione ottima) dirada i quanti in corrispondenza dei livelli meno probabili del segnale, in modo che l'errore medio globale ne risenta marginalmente; il secondo (quantizzazione logaritmica) utilizza quanti di ampiezza inferiore per i livelli più bassi del segnale e viceversa per i livelli più alti, tentando di rendere costante il rapporto segnale rumore.

4.2 QUANTIZZAZIONE OTTIMA NON UNIFORME

Fissata la soglia di saturazione V, per massimizzare l'efficienza della codifica, nel senso della minimizzazione globale dell'errore di quantizzazione, è necessario determinare opportunamente la caratteristica del quantizzatore f(x)

in funzione della distribuzione delle ampiezze del segnale p(x). A tal fine, è necessario determinare la caratteristica che minimizzi il funzionale dell'errore

$$J = \int_0^V \frac{p(x)}{f'^2(x)} dx = \int_0^V F[x, f'(x)] dx \tag{4.1}$$

Tale caratteristica deve essere trovata nella famiglia di curve $f(x) + \varepsilon \mu(x)$, ottenute tramite una variazione della f(x) ottima, nel caso di estremi vincolati $f(0) = 0$; $f(V) = V$ (cioè $\mu(0) = \mu(V) = 0$) (fig. 4.2). Per la ricerca del minimo, si calcola l'incremento del funzionale a seguito della variazione della f(x), che, nel caso più generale di $F = F[x, f(x), f'(x)]$, è dato da

$$\begin{aligned}\Delta J &= J[x, f(x) + \varepsilon \mu(x), f'(x) + \varepsilon \mu'(x)] - J[x, f(x), f'(x)] \\ &= \int_0^V F[x, f(x) + \varepsilon \mu(x), f'(x) + \varepsilon \mu'(x)] - F[x, f(x), f'(x)] \, dx\end{aligned} \tag{4.2}$$

È conveniente approssimare tale incremento tramite un suo sviluppo in serie di Taylor. A questo proposito si consideri che l'incremento ($\varepsilon\mu$ per la f(x) e $\varepsilon\mu'$ per la f'(x)) non interessa la x, per cui la derivata parziale della F relativa alla x si annulla. Lo sviluppo porta alla seguente espressione

$$\Delta J \approx \int_0^V \varepsilon \left(\frac{\partial F}{\partial f} \mu + \frac{\partial F}{\partial f'} \mu' \right) dx \tag{4.3}$$

Separando i due termini dell'integrale ed integrando per parti il secondo termine, si ottiene

$$\Delta J \approx \varepsilon \int_0^V \mu \frac{\partial F}{\partial f} dx + \left\{ \varepsilon \mu \frac{\partial F}{\partial f'} \right\}_0^V - \varepsilon \int_0^V \mu \frac{d}{dx} \left(\frac{\partial F}{\partial f'} \right) dx \tag{4.4}$$

Dato che $\mu(V) = \mu(0) = 0$, il termine centrale dell'espressione è nullo, per cui

$$\Delta J \approx \varepsilon \int_0^V \mu \left[\frac{\partial F}{\partial f} - \frac{d}{dx} \left(\frac{\partial F}{\partial f'} \right) \right] dx \tag{4.5}$$

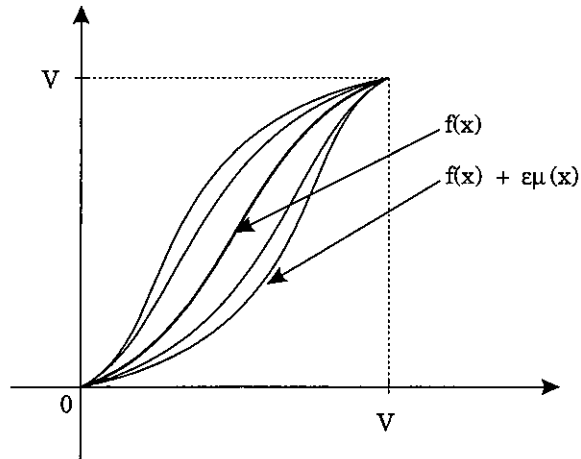



Fig. 4.2 - Variazioni della caratteristica ottima.

Nel minimo del funzionale J, l'incremento ΔJ deve annullarsi per ε che tende a zero [Gel63]. Nella nostra espressione è necessario, quindi, che si annulli l'argomento dell'integrale per qualsiasi funzione μ(x) e cioè che si annulli il termine tra parentesi quadre. L'espressione che si ricava è l'equazione di Eulero

$$\begin{cases} \frac{d}{dx}\frac{\partial F}{\partial f'} - \frac{\partial F}{\partial f} = 0 \\ f(0) = 0\,;\ f(V) = V \end{cases} \tag{4.6}$$

Nel nostro caso la F non dipende esplicitamente dalla f(x), per cui l'equazione si semplifica in

$$\begin{cases} \frac{d}{dx}\frac{\partial F}{\partial f'} = 0 \\ F = \frac{p(x)}{f'^2(x)} \end{cases} \tag{4.7}$$

Procedendo al calcolo delle derivate, si ottiene

$$\frac{\partial F}{\partial f'} = 2\,\frac{p(x)}{f'^3(x)}$$

$$\frac{d}{dx}\frac{p(x)}{f'^3(x)} = \frac{p'(x)}{f'^3(x)} - \frac{3\,p(x)\,f''(x)}{f'^4(x)} \tag{4.8}$$

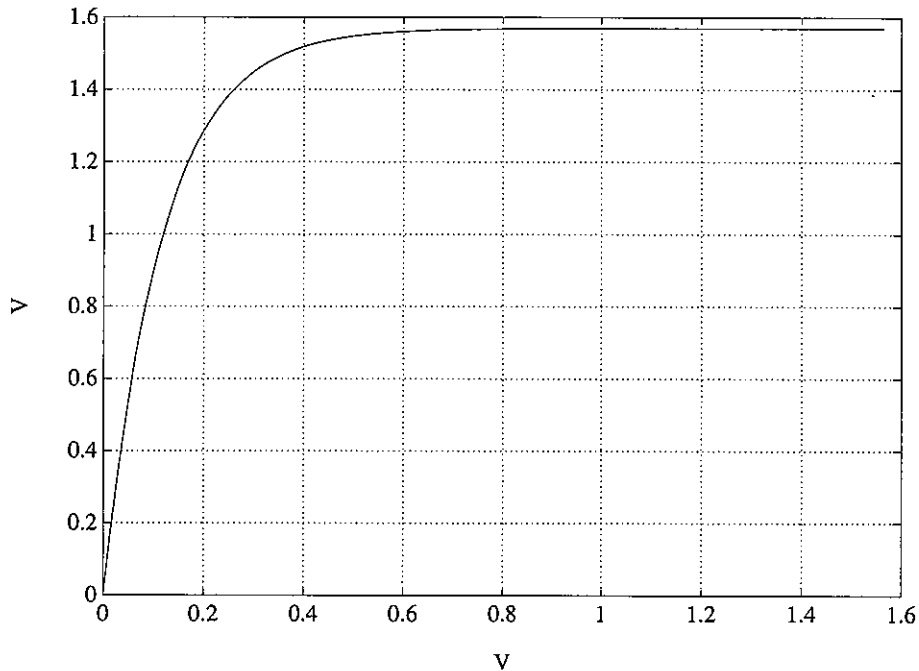



Fig. 4.3 - Caratteristica per quantizzazione ottima.

Imponendo che si annulli tale derivata si ricava

$$p'(x) = 3\,\frac{p(x)\,f''(x)}{f'(x)} \tag{4.9}$$

e separando le variabili si ottiene infine

$$\frac{p'(x)}{p(x)} = 3\,\frac{f''(x)}{f'(x)} \tag{4.10}$$

Si riconosce che entrambi i membri dell'equazione sono le derivate dei logaritmi dei denominatori, per cui questa equazione può essere integrata come

$$\log p(x) = 3 \log f'(x) + k_0 = 3 \log f'(x) + 3 \log k_1$$

$$f'(x) = \sqrt[3]{k_2\, p(x)} \tag{4.11}$$

La caratteristica f(x) dipende, quindi, dalla distribuzione di probabilità p(x) delle ampiezze del segnale. Utilizzando per la p(x) una distribuzione esponenziale, si ottiene

$$f'(x) = k_3 \, e^{\frac{\alpha}{3}|x|}$$

$$f(x) = - k_4 \, e^{-\frac{\alpha}{3}|x|} + k_5 \qquad (4.12)$$

Imponendo le condizioni a contorno f(0) = 0 e f(V) = V, si trova, infine

$$f(x) = V \, \frac{1 - e^{-\frac{\alpha}{3}x}}{1 - e^{-\frac{\alpha}{3}V}} \qquad (4.13)$$

Questa è l'espressione cercata della caratteristica che minimizza l'errore globale di quantizzazione (fig. 4.3) e la quantizzazione ottenuta secondo tale caratteristica non lineare viene definita ottima. Studiandone l'andamento per un fissato valore $\overline{\alpha}$ dell'esponente (ad esempio per il valore efficace corrispondente ad una potenza media di – 23 dBm0 [Bon91]), si osserva che essa è tale da assegnare quanti di ampiezza minore per i livelli inferiori del segnale e viceversa per i livelli superiori. Dato che il contributo maggiore all'errore globale deriva dai livelli inferiori, che sono i più probabili, in tal modo è possibile ridurre globalmente la potenza del rumore rispetto alla quantizzazione uniforme. Inoltre, alla minimizzazione dell'errore globale, si affianca in questo caso un andamento del rapporto segnale rumore vantaggioso per i livelli più bassi. Questo, però, non risulta vero in generale. Infatti, per un segnale sinusoidale, dove i livelli più probabili del segnale sono quelli maggiori, la quantizzazione ottima addenserà i quanti per tali valori delle ampiezze. Rispetto ad una quantizzazione uniforme, il rapporto segnale rumore, quindi, risulterà ulteriormente peggiorato per i livelli più bassi.

Fissata la caratteristica, si può ricavare la potenza del rumore per un segnale con un qualsiasi valore efficace. Sostituendo la f(x) nell'espressione generale della componente granulare, si ottiene

$$f'(x) = \frac{\overline{\alpha}\, V}{3} \, \frac{e^{-\frac{\overline{\alpha}}{3}x}}{1 - e^{-\frac{\overline{\alpha}}{3}V}}$$

$$e^2_{gn} = \frac{2V^2}{3n^2} \int_0^V \frac{p(x)}{f'^2(x)} \, dx = \frac{3}{\overline{\alpha}^2 \, n^2} \left(1 - e^{-\frac{\overline{\alpha}}{3}V}\right)^2 \alpha \int_0^V e^{\left(2\frac{\overline{\alpha}}{3} - \alpha\right)x} \, dx$$

$$e_{gn}^2 = \frac{9\,\alpha}{\overline{\alpha}^2\, n^2} \left(1 - e^{-\frac{\overline{\alpha}}{3} V}\right)^2 \frac{1 - e^{\left(2\frac{\overline{\alpha}}{3} - \alpha\right) V}}{3\,\alpha - 2\,\overline{\alpha}} \tag{4.14}$$

Tale quantizzazione non uniforme può essere anche interpretata come una quantizzazione uniforme eseguita su una versione compressa del segnale. Le componenti che subiscono l'effetto di saturazione introdotto dalla caratteristica, cioè quelle che contribuiscono ad accrescere la potenza del rumore, sono quelle ad ampiezza maggiore. Il contributo principale all'errore deriva, quindi, dalle "code" della distribuzione di probabilità. Se si confronta l'errore con quello ottenuto nel caso di quantizzazione uniforme al variare del valore efficace del segnale, si osserva un notevole guadagno se il valore efficace del segnale coincide con quello adottando per il calcolo della caratteristica. Utilizzando segnali con un valore efficace superiore a quello utilizzato per la il calcolo della caratteristica ottima, però, il rapporto S/N peggiora sensibilmente e tale effetto è particolarmente sentito nel caso di distribuzione esponenziale, la cui densità decresce con e^{-x}. Passando ad una distribuzione gaussiana, infatti, la cui densità decresce con e^{-x^2}, si ha un errore pari a

$$e_{gn}^2 = \frac{2V^2}{3n^2} \int_0^V \frac{p(x)}{f'^2(x)}\, dx = \frac{6}{\overline{\alpha}^2\, n^2} \left(1 - e^{-\frac{\overline{\alpha}}{3} V}\right)^2 \sqrt{\frac{\beta}{\pi}} \int_0^V e^{\frac{2}{3}\overline{\alpha} x - \beta x^2}\, dx$$

$$e_{gn}^2 = \frac{3}{\overline{\alpha}^2\, n^2} \left(1 - e^{-\frac{\overline{\alpha}}{3} V}\right)^2 \left\{ \operatorname{erf}\left[\sqrt{\beta} \left(\frac{-\overline{\alpha}}{3\,\beta} + V \right) \right] + \operatorname{erf}\left[\frac{\overline{\alpha}}{3\sqrt{\beta}} \right] \right\} e^{\frac{\overline{\alpha}^2}{9\,\beta}} \tag{4.15}$$

ed il peggioramento del rapporto S/N si osserva per valori efficaci del segnale più elevati rispetto al caso dell'esponenziale (fig. 4.4).

Anche nella quantizzazione non uniforme è possibile minimizzare l'errore globale ottimizzando contemporaneamente la componente granulare e di sovraccarico. I livelli di quantizzazione, ottenuti numericamente, nel caso di segnali con distribuzione gaussiana sono riportati in tabella 4.1 dove vengono indicati i livelli di decisione x ed il livelli di restituzione y per quantizzatori caratterizzati da differenti valori del numero di bit R [Jay84].

R	Livelli															
	1		2		3		4		5		6		7		8	
	x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y
1	0	.798														
2	0	.453	.982	1.51												
3	0	.245	.501	.756	1.05	1.34	1.75	2.15								
4	0	.128	.258	.388	.522	.657	.800	.942	1.10	1.26	1.44	1.61	1.85	2.07	2.40	2.73

Tab. 4.1 - Livelli per quantizzazione ottima non uniforme.

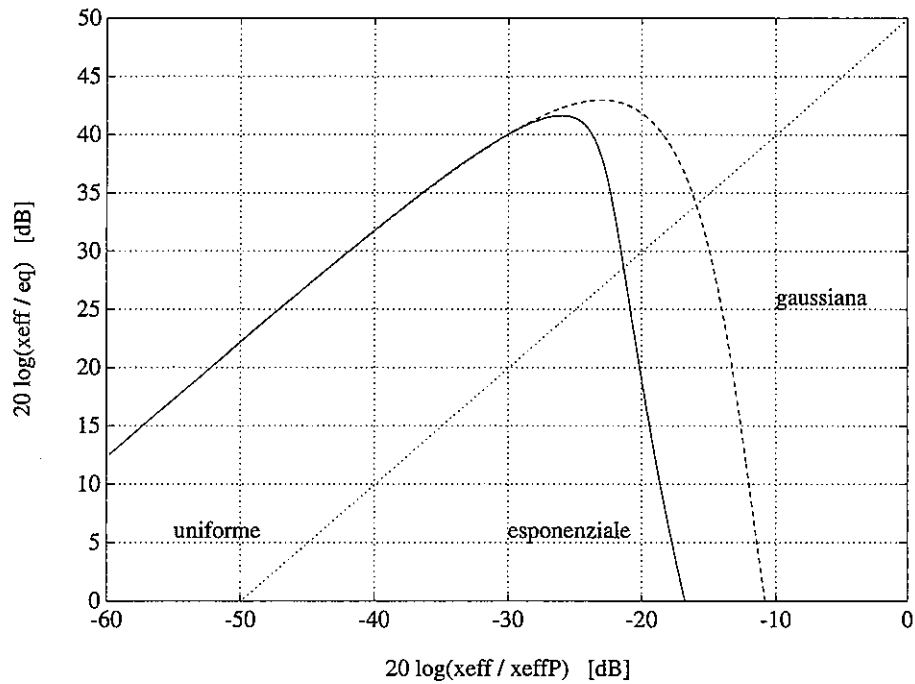



Fig. 4.4 - Rapporto S/N per quantizzazione ottima ed uniforme.

4.3 CODIFICA PCM LOGARITMICA

Risolto il problema di determinare la caratteristica che minimizza globalmente l'errore di quantizzazione, si vuole ora determinare la caratteristica in grado di rendere sufficientemente costante il rapporto segnale rumore di

quantizzazione, al variare dall'ampiezza dei campioni. Con tale criterio, oltre ad ottimizzare il rapporto S/N, non viene richiesta nessuna conoscenza sulla distribuzione delle ampiezze del segnale.

Per ottenere la f(x), è necessario, innanzitutto calcolare l'errore quadratico medio all'interno del generico quanto i-esimo, ottenibile dall'espressione generale dell'errore di quantizzazione ponendo $x = x_i$

$$e^2_{gi} = \frac{V^2}{3n^2 f'^2(x_i)} \qquad (4.16)$$

Passando al rapporto segnale-rumore, è necessario imporre che il rapporto tra tale valore e la potenza del segnale $s^2 = x_i^2$ sia costante, cioè

$$\frac{s^2}{e^2_g} = \frac{3\, x_i^2}{V^2} 2^{2R} f'^2(x_i) = k_1 \qquad (4.17)$$

Integrando per separazione di variabili, si ottiene

$$f'(x_i) = \frac{df(x_i)}{dx_i} = \frac{k_2}{x_i}$$
$$df(x_i) = k_2 \frac{dx_i}{x_i}$$
$$f(x_i) = k_2 \log(x_i) + k_3 \qquad (4.18)$$

Il legame tra ingresso ed uscita è, quindi, dato da una legge logaritmica ed il codificatore relativo è indicato come LogPCM. Dato che l'uso di tale caratteristica può essere interpretato in trasmissione come una compressione del segnale ed in ricezione come una sua espansione, la trasformazione del segnale tramite quantizzazione logaritmica è normalmente detta companding (**compression-expanding**).

Una legge puramente logaritmica non è praticamente realizzabile, dato che questa richiederebbe l'utilizzo di un numero infinito di intervalli per ampiezze prossime allo zero. Esistono due principali approssimazioni di tale logaritmica [CCITT G.711]. Una prima legge (legge m), diffusa nel mercato nordamericano e giapponese, si ottiene imponendo che la caratteristica logaritmica passi per l'origine, tramite una traslazione degli assi. Indicando con x la tensione d'ingresso normaliz-

zata rispetto a quella di saturazione V, il legame tra ingresso ed uscita è dato da

$$f(x) = \mathrm{sgn}(x) \frac{\ln(1 + \mu |x|)}{\ln(1 + \mu)} \quad -1 \le x \le 1 \tag{4.19}$$

dove $\mu = 255$, $V = 3.17$ dBm0 ed il un numero di bit utilizzati è pari a 7. Una seconda legge (legge A), diffusa nel mercato europeo, approssima la parte della caratteristica relativa ai livelli inferiori tramite una legge di quantizzazione lineare. In tal caso

$$f(x) = \begin{cases} \mathrm{sgn}(x) \dfrac{A |x|}{1 + \ln A} & 0 \le |x| \le 1/A \\ \mathrm{sgn}(x) \dfrac{1 + \ln A |x|}{1 + \ln A} & 1/A \le |x| \le 1 \end{cases} \tag{4.20}$$

con $A = 87.6$, $V = 3.14$ dBm0 ed il un numero di bit utilizzati è pari ad 8.

Analizzando l'andamento del rapporto segnale rumore con tali leggi logaritmiche, è necessario introdurre l'espressione della caratteristica nell'e-spressione dell'errore granulare

$$e_g^2 = \frac{2V^2}{3n^2} \int_0^V \frac{p(x)}{f'^2(x)} \, dx \tag{4.21}$$

Il calcolo delle derivate delle leggi μ ed A porta, rispettivamente, ai seguenti risultati

$$f'_\mu(x) = \frac{\mu}{(1 + \mu x) \ \ln (1 + \mu)}$$

$$f'_A(x) = \begin{cases} \dfrac{A}{1 + \ln A} & 0 \le x \le 1/A \\ \dfrac{1}{x \ (1 + \ln A)} & 1/A \le x \le 1 \end{cases} \tag{4.22}$$

Considerando segnali con distribuzione delle ampiezze esponenziale, l'errore è, quindi, pari a

$$e_{g\mu}^2 = \frac{\alpha V^2}{3n^2} \frac{\ln^2(1 + \mu)}{\mu^2} \int_0^V (1 + \mu x)^2 e^{-\alpha x} \, dx$$

$$e_{gA}^2 = \frac{\alpha V^2}{3n^2} (1 + \ln A)^2 \left(\int_0^{\frac{1}{A}} \frac{e^{-\alpha x}}{A^2} \, dx + \int_{\frac{1}{A}}^V x^2 e^{-\alpha x} \, dx \right) \tag{4.23}$$

dove

$$\int (1+\mu x)^2 e^{-\alpha x}\, dx = -\frac{\left(\alpha^2+2\alpha\mu+2\mu^2\right)+2\alpha\mu\,(\alpha+\mu)\,x+\alpha^2\mu^2x^2}{\alpha^3}\, e^{-\alpha x}$$

$$\int \frac{e^{-\alpha x}}{A^2}\, dx = \frac{1}{A^2\alpha}\, e^{-\alpha x}$$

$$\int x^2 e^{-\alpha x}\, dx = \frac{2+2\alpha x+\alpha^2x^2}{\alpha^3}\, e^{-\alpha x} \tag{4.24}$$

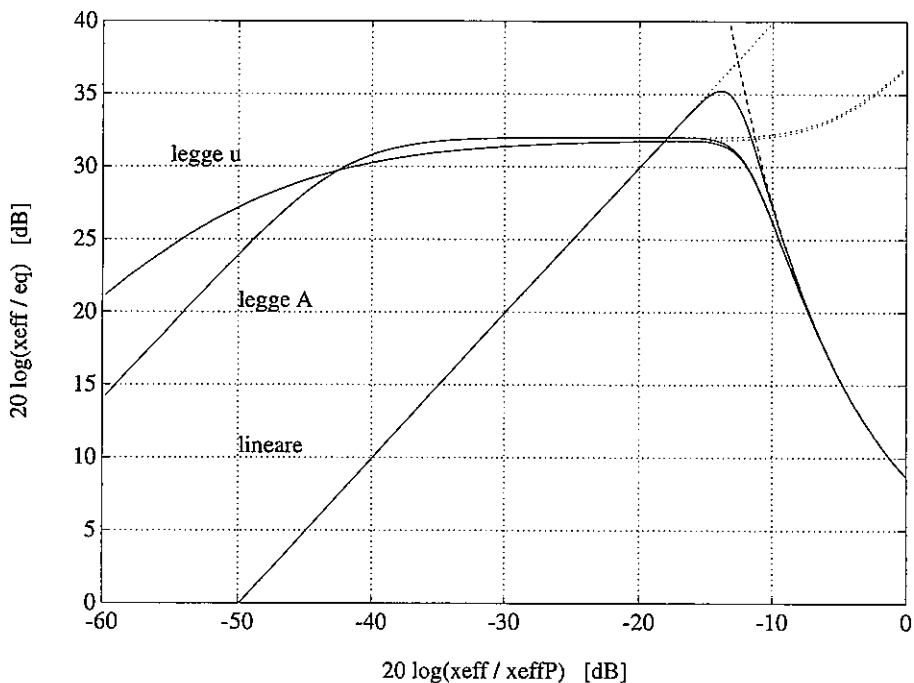



Fig. 4.5 - Rapporto segnale/rumore di quantizzazione per differenti caratteristiche.

Il rapporto segnale rumore ottenibile con tale codifica è mostrato in figura 4.5. Si nota, innanzitutto, un innalzamento della curva relativa alla quantizzazione logaritmica rispetto ad una quantizzazione lineare a parità di numero di bit. Inoltre, tale rapporto migliora linearmente con l'aumento dell'ampiezza del segnale fino a che il livello stesso rimane nell'intorno dello zero, dopo di che la curva si appiattisce a seguito dell'incremento del passo di

quantizzazione. Raggiunto poi il livello di saturazione, il rapporto segnale rumore inizia a decrescere. È opportuno chiarire che l'innalzamento della curva della componente granulare che si osserva in figura dopo la fase a rapporto segnale/rumore costante è dovuto al fatto che ormai la quasi totalità della potenza del rumore di quantizzazione è dato dalla componente di saturazione.

Considerando la quantizzazione non uniforme specificata dalla legge A, l'implementazione si basa su una conversione A/D con quantizzazione uniforme su 13 bit ed una successiva compressione numerica su 8 bit. La compressione utilizza un'approssimazione della legge logaritmica ottenuta tramite una spezzata (fig. 4.6). Questa viene fissata suddividendo la dinamica del segnale innanzitutto in due polarità (rappresentata dal bit più significativo del codice compresso). Per ciascuna polarità vengono individuati 8 segmenti, codificati con i tre bit immediatamente meno significativi. Il codice del segmento è ottenibile sottraendo a 7 il numero di zeri iniziali del campione. A ciascun segmento è associato un passo di quantizzazione che raddoppia progressivamente di ampiezza nel passaggio da un segmento all'altro, ad esclusione dei primi due segmenti, come mostrato nella tabella 4.2.

Ingresso	Passo	Codifica	Decodifica	Ampiezza
0 - 15	1	000	0 - 15	0.5 - 15.5
16 - 31	1	001	16 - 31	16.5 - 31.5
32 - 63	2	010	32 - 47	33 - 63
64 - 127	4	011	48 - 63	66 - 126
128 - 255	8	100	64 - 79	132 - 252
256 - 511	16	101	80 - 95	264 - 504
512 - 1023	32	110	96 - 111	528 - 1008
1024 - 2047	64	111	112 - 127	1056 - 2016

Tab. 4.2 - Passi di quantizzazione per LogPCM.

All'interno di ciascun segmento la codifica è lineare su 16 livelli ed è espressa negli ultimi quattro bit del codice compresso. Per quanto riguarda i primi due segmenti, dato che il quanto ad essi associato è delle stesse dimensioni di quello del quantizzatore uniforme, i 4 bit di codifica si ottengono dai 5 bit meno significativi della codifica lineare, escludendo il LSB. Agli altri segmenti è poi associato un quanto crescente con potenze di 2 man mano che ci si sposta verso valori crescenti dell'ampiezza. Ciò vuol dire che la codifica

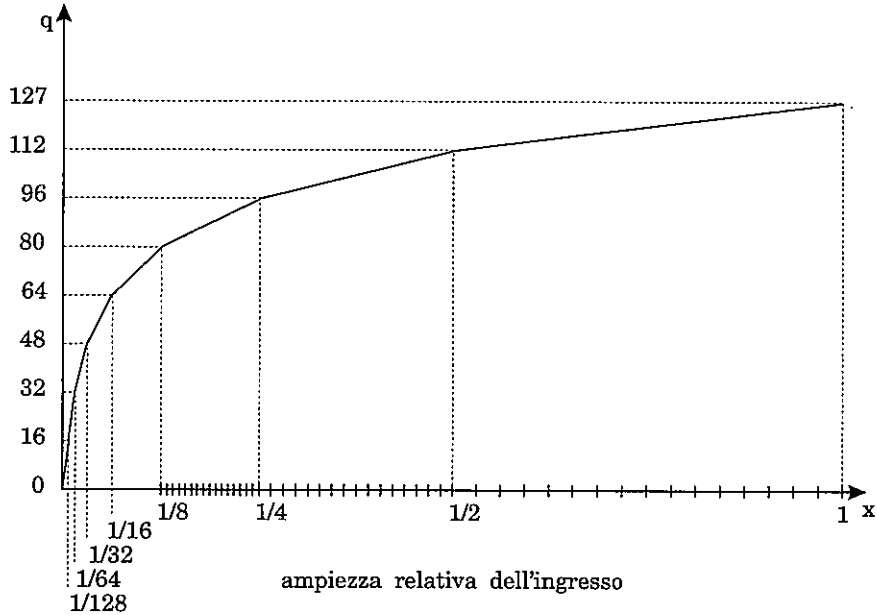



Fig. 4.6 - Approssimazione a tratti della caratteristica non lineare.

all'interno di un segmento è ottenibile utilizzando una finestra di 4 bit nella codifica lineare, in posizioni sempre più spostate verso il bit più significativo all'aumentare del numero di segmento considerato (fig. 4.7), come mostrato in tabella 4.3.

Ingresso lineare												Uscita compressa						
11	10	9	8	7	6	5	4	3	2	1	0	6	5	4	3	2	1	0
0	0	0	0	0	0	0	Q3	Q2	Q1	Q0	x	0	0	0	Q3	Q2	Q1	Q0
0	0	0	0	0	0	1	Q3	Q2	Q1	Q0	x	0	0	1	Q3	Q2	Q1	Q0
0	0	0	0	0	1	Q3	Q2	Q1	Q0	x	x	0	1	0	Q3	Q2	Q1	Q0
0	0	0	0	1	Q3	Q2	Q1	Q0	x	x	x	0	1	1	Q3	Q2	Q1	Q0
0	0	0	1	Q3	Q2	Q1	Q0	x	x	x	x	1	0	0	Q3	Q2	Q1	Q0
0	0	1	Q3	Q2	Q1	Q0	x	x	x	x	x	1	0	1	Q3	Q2	Q1	Q0
0	1	Q3	Q2	Q1	Q0	x	x	x	x	x	x	1	1	0	Q3	Q2	Q1	Q0
1	Q3	Q2	Q1	Q0	x	x	x	x	x	x	x	1	1	1	Q3	Q2	Q1	Q0

Tab. 4.3 - Legame tra codifica PCM e LogPCM.

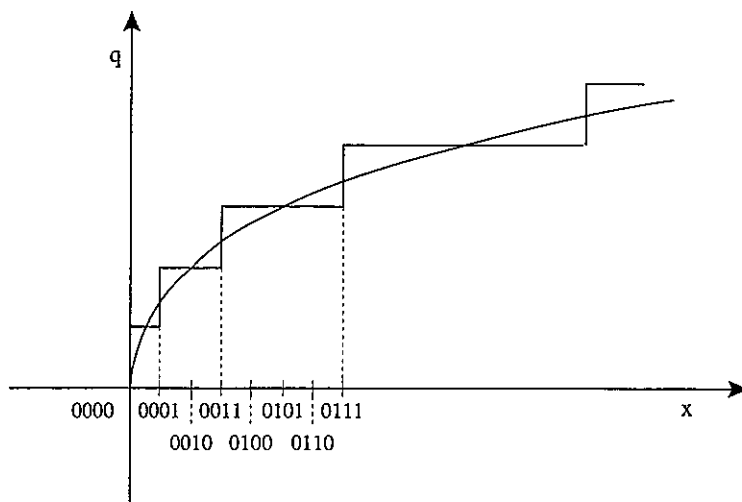



Fig. 4.7 - Perdita di bit meno significativi all'aumentare dell'ampiezza del quanto.

La procedura precedentemente esposta può essere interpretata come la trasformazione di un numero intero a 12 bit in una rappresentazione floating-point, nella quale la mantissa è su 5 bit (quantizzazione più segno), mentre l'esponente è su 3 bit (segmento). In codifica, quindi, si utilizzano solamente i 12 MSB della codifica lineare. Il LSB è utilizzato in decodifica (ponendolo pari ad 1) per posizionare il livello di restituzione a metà del quanto. Tale azione è necessaria per ridurre l'influenza del rumore per segnali prossimi allo zero.

In figura 4.8 si riportano le maschere del rapporto segnale/rumore per segnali di prova sinusoidali e con distribuzione gaussiana [ITU-T G.712]. Nel caso di segnale sinusoidale e quindi di ampiezza deterministica, si nota che l'ingresso nella zona lineare e di saturazione è più marcato di quanto avvenga con il segnale a statistica gaussiana

Data una frequenza di campionamento di 8 kHz, sfruttando codifica LogPCM è possibile esprimere ciascun campione su 8 bit invece dei 12 della quantizzazione uniforme, producendo un flusso numerico di 64 kb/s, contro i 96 kb/s originali.

segnale di prova a statistica gaussiana

segnale di prova sinusoidale

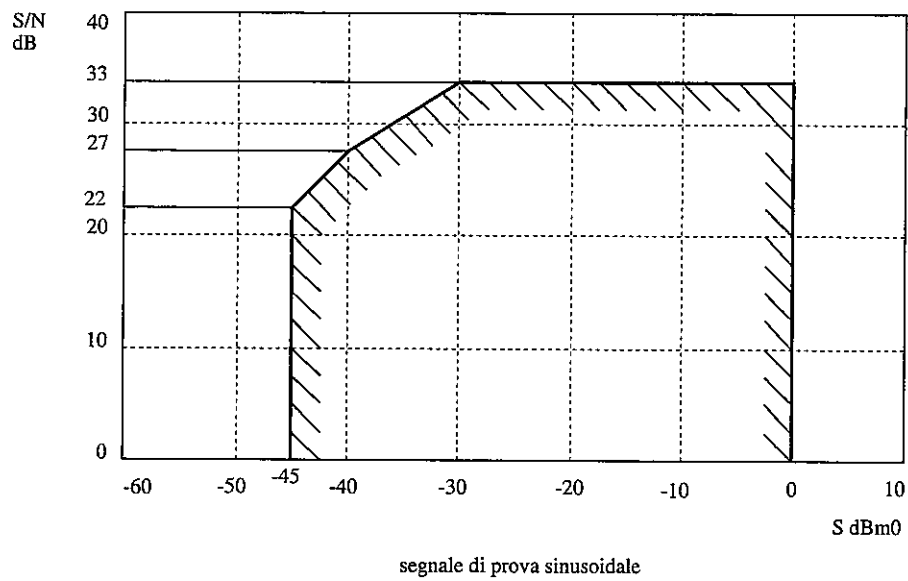
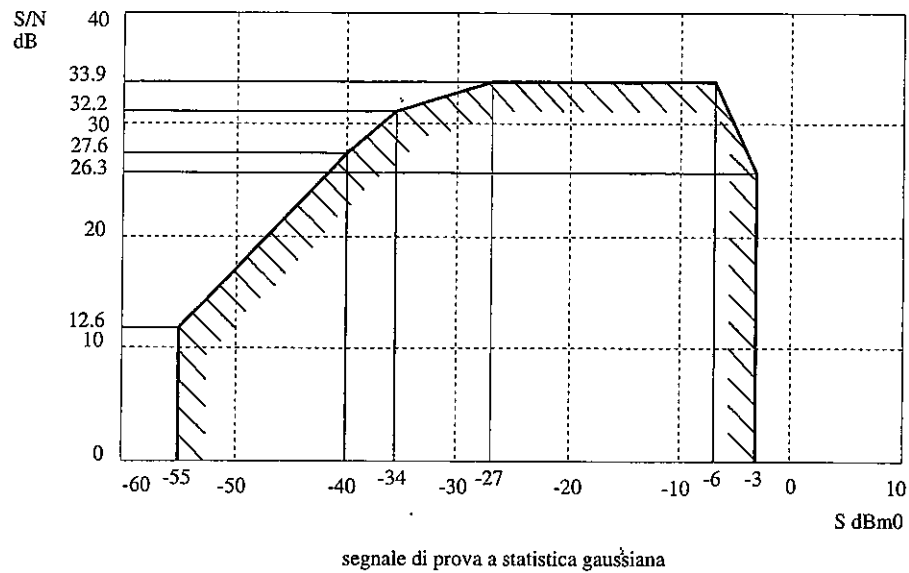



Fig. 4.8 - Maschera del rapporto segnale-rumore per quantizzatore logaritmica

5

CODIFICA NUMERICA DI FORMA D'ONDA CON MEMORIA

5.1 QUANTIZZAZIONE ADATTATIVA

Nella quantizzazione uniforme precedentemente descritta, la caratteristica di quantizzazione utilizzata è fissa nel tempo ed opera sui singoli campioni senza tener conto dell'andamento del segnale. Una simile tecnica di codifica viene indicata come "codifica di forma d'onda senza memoria" o "quantizzazione istantanea". L'ampiezza dei quanti utilizzati deriva da limiti di saturazione dimensionati per la massima dinamica del segnale, anche se essa può risultare nel tempo notevolmente inferiore.

Nel LogPCM, pur rimanendo sempre nell'ambito delle codifiche di forma d'onda senza memoria, la situazione è leggermente differente. Infatti, grazie all'approssimazione a tratti della caratteristica logaritmica, è possibile pensare la codifica LogPCM come una quantizzazione uniforme ottenuta utilizzando quella che, in un gruppo di sette distinte caratteristiche, fornisce un limite di saturazione adeguato alla dinamica del segnale (fig. 5.1). L'informazione relativa alla caratteristica utilizzata è trasmessa insieme alla quantizzazione all'interno del segmento.

In tal modo gli estremi di saturazione del quantizzatore vengono continuamente adattati alla dinamica corrente del segnale. Trasmettere ad ogni campione l'informazione relativa alla dinamica, però, equivale implicitamente ad ipotizzare che nel passaggio da un campione al successivo, l'ampiezza possa variare arbitrariamente. Nell'ipotesi di campioni correlati, però, ciò non accade ed il passo di quantizzazione può essere aggiornato con una frequenza inferiore

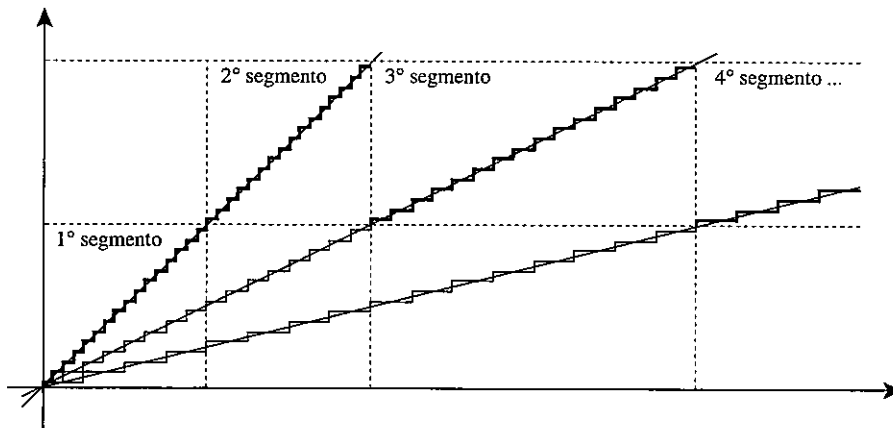



Fig. 5.1 - Approssimazione tramite quantizzazione uniforme del LogPCM.

in funzione del valore efficace dell'ingresso (fig. 5.2). Ciò permette di evitare la trasmissione dell'informazione sul segmento, riducendo il numero di bit per ogni campione richiesti dalla codifica. D'altra parte è richiesto che nella codifica si abbia memoria dell'andamento del segnale, per cui la quantizzazione adattativa rientra nella famiglia delle "codifiche di forma d'onda con memoria."

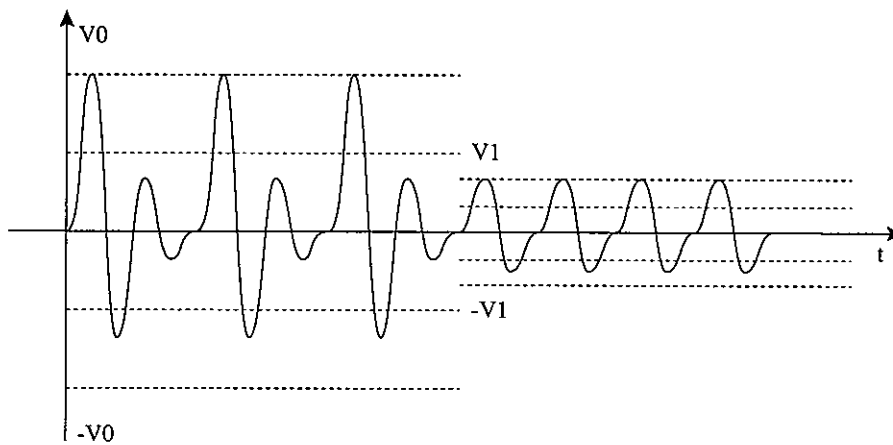



Fig. 5.2 - Quantizzazione adattativa.

Il problema è quindi quello di identificare quale caratteristica uniforme adottare (e cioè definire il numero di quanti) e fissare la legge con la quale adattare l'ampiezza del quanto (e quindi cambiare tipo di caratteristica). La definizione del numero di quanti può essere eseguita in funzione del rumore di quantizzazione ammissibile. Volendo mantenere le prestazioni ottenute da un codificatore LogPCM, è necessario determinare il numero di bit utilizzati da questo nella quantizzazione uniforme all'interno di ciascun segmento.

Grazie al bit di segno ed ai quattro bit di mantissa, nel LogPCM si identifica uno tra i 16 possibili quanti presenti nell'intervallo di ampiezze compreso tra il limite di saturazione e metà della dinamica di ciascun quantizzatore. Per coprire l'intera dinamica è, quindi, necessario aggiungere un solo bit alla codifica. Il numero di bit richiesto per un quantizzatore uniforme che produca lo stesso rumore di quantizzazione del LogPCM è, quindi, pari a 6.

Per quanto riguarda l'aggiornamento del passo di quantizzazione e quindi la variazione del tipo di caratteristica, questo si può pensare come ottenuto dalla cascata di un amplificatore con controllo automatico di guadagno e di un quantizzatore uniforme con livelli di saturazione (e, di conseguenza, ampiezze dei quanti) fissi. Per l'aggiornamento del passo di quantizzazione, approssimando l'energia a breve termine con la varianza del segnale σ_x^2 e considerando un guadagno proporzionale alla varianza stessa, si può adottare una legge del tipo

$$\begin{cases} \Delta(n) = \sigma_x(n)\ \Delta_0 \\ q(n) = \dfrac{x(n)}{\Delta(n)} \end{cases} \tag{5.1}$$

dove q(n) sono le uscite del quantizzatore e Δ_0 è una costante tale che i valori di $\Delta(n)$ varino tra minimi e massimi pari a quelli utilizzati nel LogPCM. Anche in questo caso, come nel LogPCM, è preferibile realizzare un sistema digitale. Per tale motivo l'algoritmo di adattamento dell'ampiezza del quanto viene posto a valle di una conversione A/D eseguita tramite una quantizzazione uniforme che, nel segnale telefonico, è su 12 bit. Di conseguenza, anche se l'algoritmo di aggiornamento porta a valori reali del $\Delta(n)$, il passo di quantizzazione da adottare non può che essere un multiplo (secondo potenze di due) del quanto utilizzato nella quantizzazione uniforme. La riquantizzazione, quindi, si traduce in eventuali traslazioni di una finestra di 6

bit sulla codifica d'ingresso, con perdita di bit meno significativi (variazione dell'ampiezza del quanto) o più significativi (variazione del limite di saturazione (fig. 5.3).

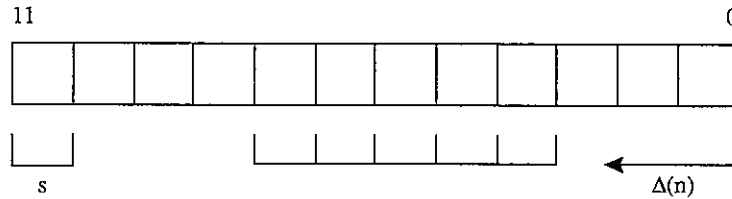



Fig. 5.3 - Conversione tra quantizzazione uniforme ed adattativa.

Il problema, quindi, è quello della stima $\hat{\sigma}_x$ del valore efficace del segnale a partire da un numero finito di campioni. Si possono seguire due metodi (fig. 5.4). Nel primo (quantizzazione adattativa in avanti o "forward": AQF) il calcolo del valore efficace viene eseguito sul blocco di campioni da codificare. Il blocco deve essere di dimensioni tali da mantenere l'ipotesi di stazionarietà della sorgente. Nel secondo metodo (quantizzazione adattativa all'indietro o "backward": AQB) il valore efficace viene calcolato campione per campione in funzione dei campioni più recenti.

Con l'adattamento in avanti, considerando blocchi di N campioni, la stima della varianza può essere ottenuta come

$$\hat{\sigma}_x^2(n) = \frac{1}{N} \sum_{m=0}^{N-1} x^2(n+m) \tag{5.2}$$

Il passo di quantizzazione che ne deriva viene utilizzato per la codifica e deve essere trasmesso al destinatario per la decodifica. Anche se questo metodo ha una buona efficienza, presenta lo svantaggio di dover trasmettere (sia pur con una banda ridotta, ma con il rischio di errori) il passo di quantizzazione. Inoltre, la necessità di accumulare un buffer di campioni, introduce un ritardo ineliminabile di codifica, che può produrre effetti indesiderati (es.: eco) in applicazioni in tempo reale.

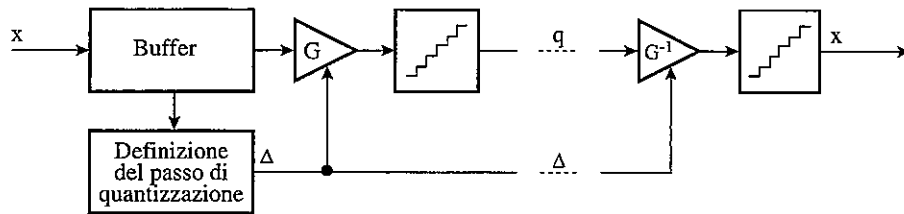
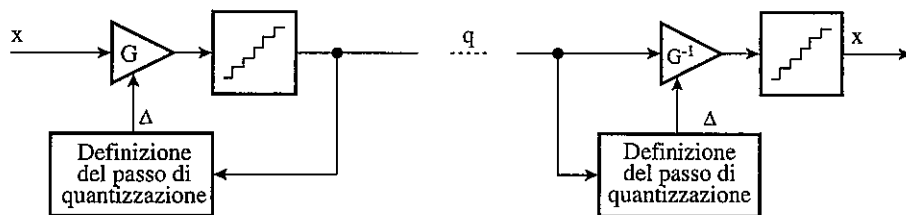
Con l'adattamento all'indietro, invece, le grandezze su cui calcolare il quanto sono disponibili sia al trasmittente che al destinatario; in questo modo

Quantizzazione adattativa in avanti

Quantizzazione adattativa all'indietro

Fig. 5.4 - Quantizzazione adattativa in avanti e all'indietro.

non è richiesta la trasmissione di nessuna informazione aggiuntiva. D'altra parte, il calcolo del valore efficace sui campioni già codificati può non risultare, ovviamente, ottimale per la codifica della rimanente parte del segnale, tanto più che l'uscita del codificatore è deteriorato dal rumore di quantizzazione. I risultati ottenibili con l'adattamento all'indietro sono, quindi, tipicamente peggiori di quelli ottenuti con l'adattamento in avanti.

Rimanendo nell'adattamento all'indietro, per il calcolo della $\hat{\sigma}_x$ è necessario, innanzitutto, definire il blocco di campioni da utilizzare. Se questo blocco venisse definito tramite una finestra rettangolare del tipo

$$\hat{\sigma}_x^2(n) = \frac{1}{N} \sum_{m=1}^{N} x^2(n-m) \tag{5.3}$$

si assegnerebbe uno stesso peso sia ai campioni più lontani che quelli più recenti del segnale. Utilizzando, invece, una finestra esponenziale

$$\hat{\sigma}_x^2(n) = (1-\alpha) \sum_{m=1}^{\infty} \alpha^{m-1} x^2(n-m) \; ; \quad 0 < \alpha < 1 \tag{5.4}$$

tramite la costante α è possibile sia determinare l'ampiezza dell'intervallo di campioni utili, sia diminuire l'influenza di quelli più lontani. Il calcolo della stima $\hat{\sigma}_x$ secondo la relazione mostrata richiederebbe la memorizzazione di un blocco di uscite del quantizzatore. L'informazione sulla storia del segnale, però, è già contenuta nella $\hat{\sigma}_x$. Infatti, estraendo il termine per m = 1, l'equazione precedente può essere riscritta in maniera ricorsiva come

$$\hat{\sigma}_x^2(n) = (1-\alpha)\ x^2(n-1) + \alpha \left\{ (1-\alpha) \sum_{k=1}^{\infty} \alpha^{k-2}\ x^2[n-k-1] \right\}$$

$$\hat{\sigma}_x^2(n) = (1-\alpha)\, x^2(n-1) + \alpha\, \hat{\sigma}_x^2(n-1) \tag{5.5}$$

Al variare di α cambiano le caratteristiche del quantizzatore (fig. 5.5). Considerando un campionamento a 8 kHz, per $\alpha = 0.99$ vengono utilizzati circa 100 campioni del segnale, corrispondenti a 12.5 ms: tali sistemi vengono quindi definiti sillabici. Per $\alpha = 0.9$ la finestra include un numero di campioni inferiore a dieci, corrispondenti a 1ms, per cui il sistema viene detto istantaneo. I sistemi sillabici, mediando su di un numero maggiore di campioni, non permettono di seguire la dinamica istantanea del segnale, ma risultano poco sensibili agli errori di trasmissione. I sistemi istantanei, invece, permettono di seguire meglio l'andamento della forma d'onda del segnale, ma risultano meno robusti nei confronti degli errori. Inoltre, nei sistemi istantanei, il passo di quantizzazione viene rapidamente ridotto nelle pause del segnale. Ciò si traduce nel rischio di saturazione alla successiva ripresa di attività della sorgente.

L'adattamento all'indietro può essere ulteriormente semplificato. Dato che $\Delta(n) = \hat{\sigma}_x^2(n)\, \Delta_0$, se si calcola il rapporto tra due quanti successivi utilizzando la relazione ricorsiva precedentemente trovata per la $\hat{\sigma}_x$, si ottiene

$$\frac{\Delta(n)}{\Delta(n-1)} = \sqrt{\frac{\hat{\sigma}_x^2(n)}{\hat{\sigma}_x^2(n-1)}} = \sqrt{\alpha + \frac{1-\alpha}{\hat{\sigma}_x^2(n-1)}\, q^2(n-1)} \tag{5.6}$$

Questa relazione può essere interpretata come un moltiplicatore M, funzione della precedente uscita del quantizzatore, della varianza del segnale e dell'esponente α, in grado di fornire il quanto corrente a partire dal precedente

$$\Delta(n) = M[\,|\,q(n-1)\,|,\, \hat{\sigma}_x^2,\, \alpha\,]\, \Delta(n-1) \tag{5.7}$$

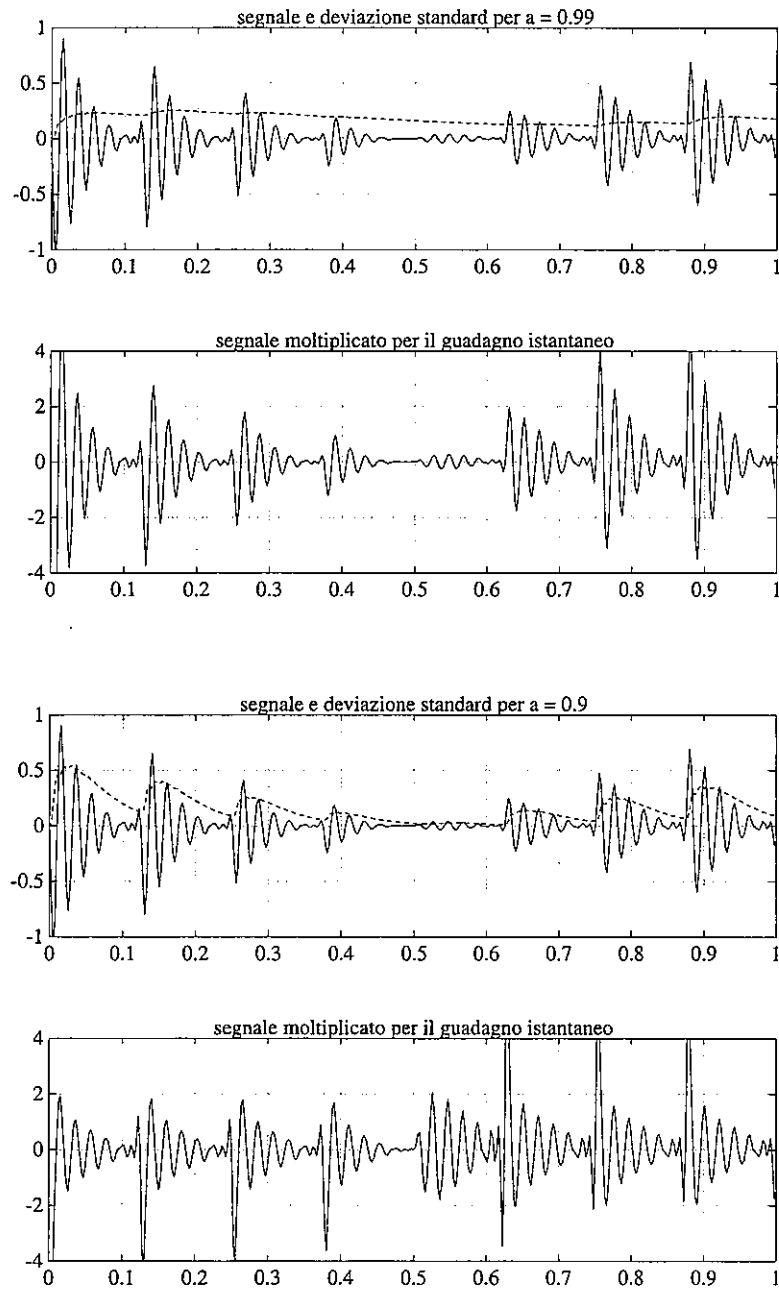



Fig. 5.5 - Segnali per sistemi sillabici ed istantanei.

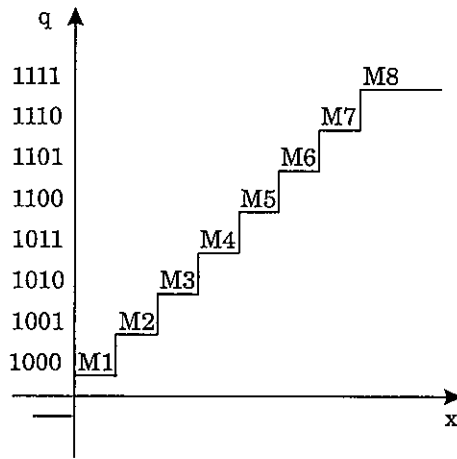



Fig. 5.6 - Legame tra ampiezza dei campioni e moltiplicatori.

L'espressione del moltiplicatore può essere semplificata, scollegandola dall'utilizzo di una finestra esponenziale, ma adottando dei moltiplicatori M costanti, funzioni della sola uscita corrente del quantizzatore (fig. 5.6). La legge di aggiornamento del passo di quantizzazione diventa del tipo

$$\Delta(n) = M[\,|q(n-1)|\,]\,\Delta(n-1) \tag{5.8}$$

La memoria dell'algoritmo in questo modo è ridotta ad un solo campione. Le tecniche per l'individuazione della funzione M(q) ottima per l'aggiornamento del passo di quantizzazione sono state a lungo analizzate [Jay76], ma i risultati mostrano che il valore adottati per la M non sembrano essere critici [Rab78] (fig. 5.7). Risulta importante, invece, che l'aumento del passo di quantizzazione sia energico man mano che il valore efficace del segnale si avvicina alla soglia di saturazione, mentre la sua riduzione in corrispondenza dei livelli più bassi di segnale può essere molto più graduale.

Una possibile scelta per i moltiplicatori di un quantizzatore adattativo ad otto livelli è riportata in tabella [ITU-T G.721]

	8	7	6	5	4	3	2	1
M_i	4.482	1.585	1.283	1.142	1.070	1.037	1.006	0.984

Tab. 5.1 - Moltiplicatori per quantizzazione adattativa.

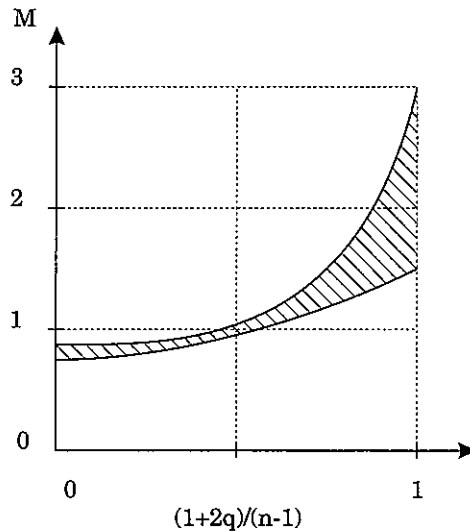



Fig. 5.7 - Area di variazione per moltiplicatori.

Si noti come il Δ venga ridotto solamente nel caso in cui il segnale cada nel primo quanto e che la riduzione è del 1.6 %. Nel caso in cui il segnale abbia raggiunto il quanto più elevato e quindi si rischi di entrare in saturazione, il quanto stesso viene aumentato più del 300 %.

Dato che con l'adozione di moltiplicatori costanti si perde il decadimento esponenziale della memoria del codificatore, torna il problema della propagazione di errori di trasmissione. È possibile rendere più robusto un quantizzatore adattativo che utilizzi dei moltiplicatori costanti agli errori di trasmissione, reintroducendo un decadimento esponenziale (robust adaptation) con una legge di aggiornamento del tipo

$$\Delta(n) = M(|q(n-1)|)\ \Delta^{1-\beta}(n-1) \qquad (5.9)$$

dove il coefficiente β (tipicamente pari a 1/32 o 1/64) smorza il contributo dei precedenti passi di quantizzazione (potenzialmente errati).

Se la quantizzazione adattativa viene applicata a campioni ottenuti da una conversione A/D lineare, la codifica che ne deriva è detta Adaptive PCM (APCM). In realtà i benefici maggiori della quantizzazione adattativa si hanno applicandola alla codifica del segnale d'errore per codificatori predittivi (ADPCM, ADM), come descritto nel seguito.

5.2 QUANTIZZAZIONE DIFFERENZIALE

5.2.1 Generalità sulla codifica predittiva

L'errore di codifica per un quantizzatore istantaneo, fissato il numero di bit, è legato al valore V degli estremi del quantizzatore. Questo, a sua volta, è funzione della dinamica del segnale. Riducendo tale dinamica, quindi, sarebbe possibile ridurre il rumore di quantizzazione o, a parità di questo, ridurre il numero di bit utilizzati nella codifica. Per procedere su tale strada si osserva qualitativamente che campioni adiacenti del segnale hanno ampiezze che si discostano tipicamente meno della massima dinamica del segnale: normalmente non si passa, cioè, dal massimo valore positivo al massimo negativo nell'intervallo di tempo che trascorre tra un campione e l'altro. Se questa ipotesi fosse vera, la codifica della differenza tra campioni adiacenti

$$d(n) = x(n) - x(n-1) \qquad (5.10)$$

porterebbe alla voluta riduzione di dinamica. Trasmettendo tale differenza, a partire da una condizione iniziale di x(0) ad esempio nulla, il segnale potrebbe essere ricostruito a destinazione sommando la differenza ricevuta dal trasmittente all'ampiezza del campione precedentemente ricostruito

$$x(n) = d(n) + x(n-1) \qquad (5.11)$$

Per procedere su tale strada è necessario verificare che la dinamica del segnale differenza sia inferiore a quella del segnale. Una grandezza dalla quale stimare la rapidità di variazione dell'ampiezza dei campioni del segnale (supposto stazionario) è l'autocorrelazione, definita come

$$R(n) = E\left\{ x(i)\, x(i+n) \right\} = \sum_{i=-\infty}^{\infty} x(i)\, x(i+n)\, p[x(i), x(i+n)] \qquad (5.12)$$

dove E{ } rappresenta l'operatore "valore atteso" e l'argomento sono copie traslate del segnale. Questa funzione, ha il suo massimo (unico) nell'origine, che è pari alla potenza del segnale

$$R(0) = E\left\{ x(i)\, x(i) \right\} = E\left\{ x^2(i) \right\} \qquad (5.13)$$

Rinviando per il momento una trattazione formale, qualitativamente si osserva che passare dal punto R(0) al punto R(n) vuol dire passare dal considerare il prodotto di due copie identiche x(i) del segnale al prodotto di due copie traslate di n campioni. Se il valore dell'autocorrelazione varia "poco," vuol dire che la x(i + n) è in media ancora "sufficientemente" simile alla x(i): l'ampiezza di campioni distanti n intervalli di campionamento, cioè, non varia "sensibilmente." Il segnale vocale, ad esempio, è un segnale che soddisfa tale condizione. Se si osserva, infatti, l'andamento della funzione di autocorrelazione a lungo termine per un segnale vocale filtrato in banda telefonica (fig. 5.8) si nota come essa abbia un andamento gradatamente decrescente, per cui la codifica della differenza delle ampiezze di campioni adiacenti risulterebbe vantaggiosa, nel senso della riduzione della dinamica.

La dinamica del segnale d'errore, però, può essere ulteriormente ridotta (fig. 5.9). L'andamento regolare della funzione di auto-correlazione, infatti, è indice della non indipendenza tra campioni del segnale. Per segnali puramente casuali a media nulla, infatti, la funzione di autocorrelazione risulterebbe essere una delta, dato che solo facendo perfettamente coincidere due copie del segnale, tutti i campioni contribuirebbero in maniera concorde nel prodotto (quindi $R(0) \neq 0$), mentre negli altri casi, essendo ciascun campione indipendente dagli altri, la risultante tenderebbe ad annullarsi. Qualsiasi forma della funzione di autocorrelazione differente dalla delta, quindi, evidenzia che i campioni sono frutto di una qualche legge di generazione.

Nel caso della voce, in particolare, essendo i campioni tutti prodotti dalla stessa sorgente tramite la propagazione dell'eccitazione nel cavo orale, è immaginabile che essi siano legati dall'andamento della risposta impulsiva di quest'ultimo. Nota la legge che lega i campioni di un segnale, è possibile ricavare una loro stima a partire dal valore dei campioni che li hanno preceduti. Maggiore è l'accuratezza della stima, minore sarà l'ampiezza del segnale d'errore, che è l'obiettivo che ci si era prefissi. Le tecniche per ottenere la stima del segnale (predizione) sono trattate nei seguenti paragrafi.

5.2.2 Richiami di predizione lineare

Il problema della stima di un segnale tramite predizione lineare rientra nel problema più generale del filtraggio adattativo, nel quale, dato un ingresso v(n), si vuol determinare la struttura di un modello discreto di una sorgente in

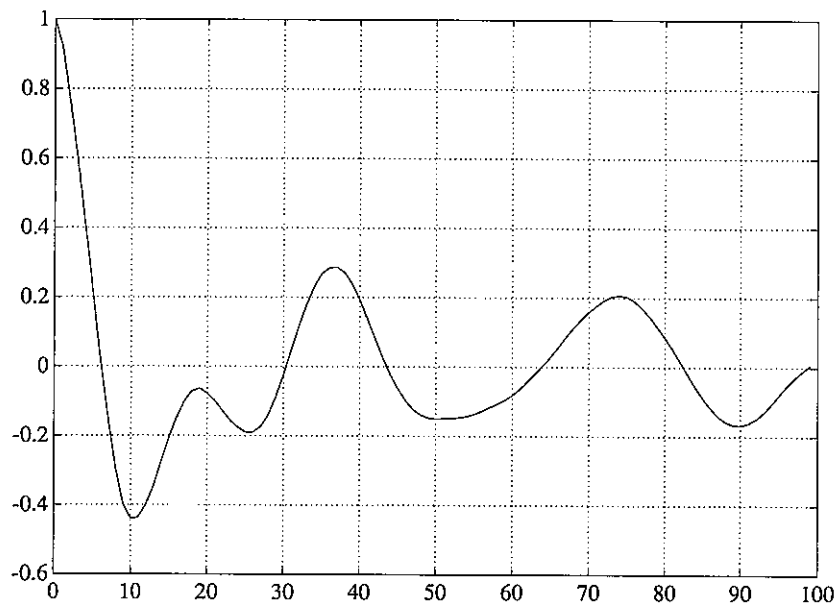



Fig. 5.8 - Funzione di autocorrelazione per segnale filtrato passa banda.

grado di produrre un'uscita desiderata u(n). Considerando segnali non deterministici, il problema è, quindi, quello di trovare un sistema in grado di generare il processo correlato u(n) a partire da un processo scorrelato v(n). Ciò è la controparte di quanto avviene per segnali deterministici con l'analisi di Fourier, nella quale un qualsiasi segnale è descritto in termini di segnali elementari sinusoidali. Nel definire il modello della sorgente, una scelta comune è quella di considerare il segnale u(n) come un processo Auto Regressivo (AR) di ordine q (fig. 5.10). Questo è l'uscita di un sistema discreto di tipo IIR, con legame ingresso/uscita esprimibile come

$$u(n) = v(n) - \sum_{k=1}^{q} w_k \, u(n-k) \tag{5.14}$$

La sua funzione di trasferimento a soli poli è pari a

$$H(z) = \frac{1}{1 + \sum_{k=1}^{q} w_k \, z^{-k}} \tag{5.15}$$

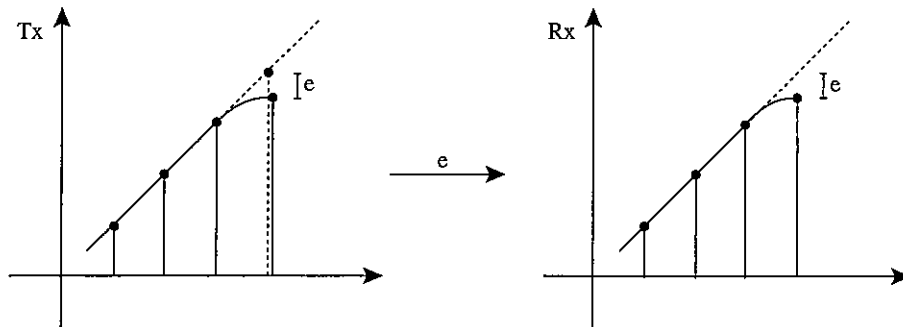



Fig. 5.9 - Codifica differenziale.

e, dato l'ingresso a spettro piatto, lo spettro dell'uscita presenta dei massimi relativi in corrispondenza alla pulsazione dei poli della H(z). L'ampiezza e la banda di tali massimi risultano funzione del modulo dei poli stessi [Appendice A.1]. Approssimare il processo correlato da analizzare con un processo AR, e quindi lineare, è essenzialmente dovuto all'esistenza di efficienti algoritmi per la soluzione del problema del filtraggio. Inoltre, per la decomposizione di Wold [Hay86], si può dimostrare che ogni processo stocastico discreto stazionario x(n) può essere scomposto nella somma di due processi scorrelati u(n) ed s(n) dove s(n) è un processo predicibile e u(n) è processo lineare (FIR con un numero infinito di termini)

$$u(n) = v(n) + \sum_{k=1}^{\infty} b_k \, v(n-k) \tag{5.16}$$

L'ingresso v(n) del FIR è sempre considerato rumore bianco. Trasformando il processo FIR in un'equivalente processo IIR (stessa risposta impulsiva) e trascurando la componente deterministica s(n), qualsiasi x(n) può essere approssimato da un processo AR.

Dato un processo AR, è possibile fissarne i coefficienti in modo tale che generi un segnale con una funzione di autocorrelazione voluta (fig. 5.11). Questo può essere facilmente provato calcolando l'autocorrelazione di un processo AR. Dall'equazione di definizione, riscritta come

$$v(n) = \sum_{k=0}^{q} w_k \, u(n-k) \, ; \; w_0 = 1 \tag{5.17}$$

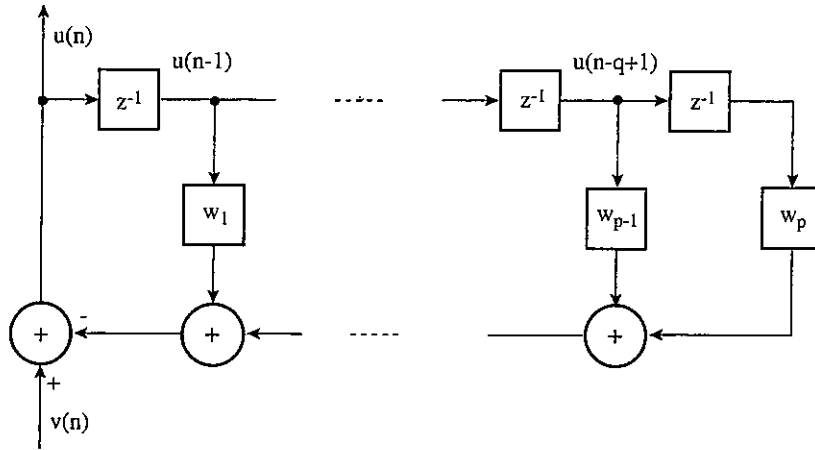



Fig. 5.10 - Rappresentazione della sorgente di un processo correlato tramite un modello Auto-Regressivo.

moltiplicando entrambi i membri per i campioni u(n i) dell'uscita ed applicando l'operatore valore atteso E{ }

$$E\{\, v(n)\, u(n-i)\,\} = E\left\{ \sum_{k=0}^{q} w_k\, u(n-k)\, u(n-i) \right\}; \quad i > 0$$

$$E\{\, v(n)\, u(n-i)\,\} = \sum_{k=0}^{q} w_k\, E\{\, u(n-k)\, u(n-i)\,\} \tag{5.18}$$

Al secondo termine si ha l'auto-correlazione R_{uu} di u(n). Il primo termine è nullo in quanto u(n - i) utilizza campioni dell'ingresso precedenti al campione v(n). Essendo quest'ultimo rumore bianco, v(n) e gli u(n-i) risultano tra loro scorrelati. Si ottiene quindi

$$\sum_{k=0}^{q} w_k\, R_{uu}(n-k) = 0\,; \quad \begin{cases} n = 1, 2, \ldots, q \\ w_0 = 1 \end{cases} \tag{5.19}$$

Esplicitando tale relazione si ottiene

$$R_{uu}(n) = \eta_1\, R_{uu}(n-1) + \eta_2\, R_{uu}(n-2) + \ldots + \eta_q\, R_{uu}(n-q)\,; \quad \eta = -\,w \tag{5.20}$$

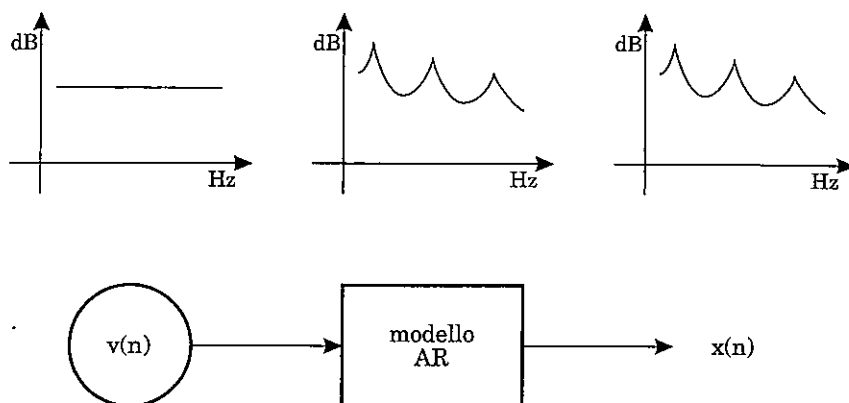



Fig. 5.11 - Relazione tra spettro del segnale e funzione di trasferimento del modello AR della sorgente.

Considerando l'indice n nell'intervallo [1, q], tale equazione può essere espressa in forma matriciale

$$\begin{bmatrix} R_{uu}(0) & R_{uu}(1) & \dots & R_{uu}(q-1) \\ R_{uu}(1) & R_{uu}(0) & \dots & R_{uu}(q-2) \\ \dots & \dots & \dots & \dots \\ R_{uu}(q-1) & R_{uu}(q-2) & \dots & R_{uu}(0) \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_q \end{bmatrix} = \begin{bmatrix} R_{uu}(1) \\ R_{uu}(2) \\ \dots \\ R_{uu}(q) \end{bmatrix} \tag{5.21}$$

ottenendo l'equazione di Yule-Walker

$$\mathbf{R_{uu}}\, \eta = \mathbf{r_{uu}} \tag{5.22}$$

Nel seguito i vettori verranno identificati in grassetto con lettere minuscole, mentre le matrici in grassetto con lettere maiuscole. Per determinare i coefficienti di un processo con autocorrelazione fissata, quindi, basta risolvere tale equazione per η

$$\eta = \mathbf{R_{uu}^{-1}}\, \mathbf{r_{uu}} \tag{5.23}$$

e sostituire i desiderati valori della $\mathbf{R_{uu}}$ [Appendice A]. Nella codifica, però, non si è interessati a generare processi con determinate caratteristiche medie (es.: autocorrelazione), ma si vuole generare un processo identico al segnale da

codificare. Per risolvere tale problema, è necessario introdurre il conceto di predizione.

La predizione lineare è un caso particolare di filtraggio nel quale il segnale desiderato coincide con l'ingresso traslato nel tempo. L'obiettivo è la determinazione della struttura di un sistema (predittore) che, in funzione di campioni disponibili del segnale, riesca a stimare il valore di campioni incogniti (fig. 5.12). In particolare, un predittore in avanti di ordine "p" stima il campione $\hat{x}$ del segnale all'istante "n" in funzione dei precedenti "p" campioni dello stesso. A tal fine si utilizza un sistema FIR con legame ingresso uscita lineare del tipo

$$\hat{x}(n) = \sum_{k=1}^{p} \alpha_k \, x(n-k) \qquad (5.24)$$

la cui funzione di trasferimento è

$$P(z) = \sum_{k=1}^{p} \alpha_k \, z^{-k} \qquad (5.25)$$

ed i coefficienti α_k sono le incognite da determinare. Anche in questo caso, l'uso di un sistema lineare è motivato da aspetti algoritmici. In forma matriciale, l'equazione del predittore diventa

$$\hat{x}(n) = [\ \alpha_1\ \alpha_2 \ldots \alpha_p\ ] \begin{bmatrix} x(n-1) \\ x(n-2) \\ \ldots \\ x(n-p) \end{bmatrix} = \alpha^T \mathbf{x}(n-1) \qquad (5.26)$$

dove i vettori

$$\alpha^T = [\ \alpha_1\ \alpha_2 \ldots \alpha_p\ ]; \quad \mathbf{x}^T(n-1) = [\ x(n-1)\ x(n-2) \ldots x(n-p)\ ] \qquad (5.27)$$

rappresentano il vettore dei coefficienti del predittore (di dimensioni pari a p) e l'ingresso dello stesso, ottenuto prelevando p elementi precedenti il campione x(n) dal vettore $\mathbf{x}(n)$ del segnale

$$\mathbf{x}^T(n) = [\ x(n)\ x(n-1) \ldots x(n-p)\ ] \qquad (5.28)$$

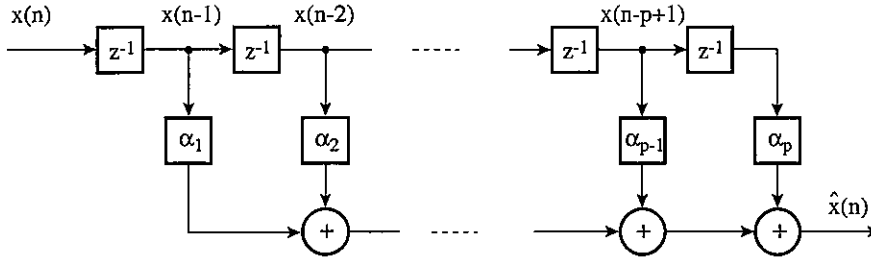



Fig. 5.12 - Predittore lineare.

L'esponente "T" indica l'operazione di trasposizione. La struttura del predittore è, ovviamente, legata a quella del processo AR che vuole stimare, come mostrato nel seguito. Dati il segnale e la sua stima, è possibile definire il vettore **e**(n) dei campioni della funzione d'errore (residuo), i cui elementi sono ottenuti come

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} \alpha_k \, x(n-k) = \mathbf{x}(n) - \alpha^T \, \mathbf{x}(n-1) \tag{5.29}$$

Dal punto di vista della predizione, l'errore è una funzione dei coefficienti incogniti α_k della quale si vuole determinare il minimo. Per determinare i parametri del predittore che minimizzino l'errore, è necessario scegliere un opportuno criterio di ottimizzazione. Quello più comunemente adottato è la minimizzazione dell'errore quadratico medio (Mean-Square value of the estimation Error: MSE), definito come la varianza del segnale d'errore

$$\varepsilon(\alpha) = E\left\{ e(n)^2 \right\} = E\left\{ [\, x(n) - \hat{x}(n) \,]^2 \right\} = E\left\{ \left[x(n) - \sum_{k=1}^{p} \alpha_k \, x(n-k) \right]^2 \right\}$$

$$\varepsilon(\alpha) = E\left\{ x(n)^2 \right\} - 2\,E\left\{ \sum_{k=1}^{p} \alpha_k \, x(n)\, x(n-k) \right\} + E\left\{ \left[\sum_{k=1}^{p} \alpha_k \, x(n-k) \right]^2 \right\} \tag{5.30}$$

Ricordando che l'autocorrelazione del segnale discreto, supposto stazionario, è definita come

$$R(n) = E\{x(i)\, x(i+n)\} = \sum_{i=-\infty}^{\infty} x(i)\, x(i+n)\, p[x(i), x(i+n)] \tag{5.31}$$

l'equazione precedente, in forma matriciale, diventa

$$\varepsilon(\alpha) = \sigma_x^2 - 2\alpha^T \mathbf{r} + \alpha^T \mathbf{R}\, \alpha \tag{5.32}$$

dove $\sigma_x^2 = R(0)$ rappresenta la varianza del segnale e

$$\mathbf{R} = \begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix}; \quad \mathbf{r} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \dots \\ R(p) \end{bmatrix} \tag{5.33}$$

sono, rispettivamente, la matrice R di autocorrelazione dell'ingresso x(n-1) (che coincide con quella di x(n)) ed il vettore di cross-correlazione tra l'ingresso x(n-1) e l'uscita desiderata x(n). Se R è una matrice definita positiva (cioè $\mathbf{y}^T \mathbf{R}\, \mathbf{y} > 0$, con y vettore arbitrario), ipotesi verificata in pratica, l'MSE è una funzione quadratica in α_k ed il suo minimo è unico (fig. 5.13). Esso si ottiene annullando contemporaneamente le derivate parziali dell'MSE rispetto al coefficiente generico α_i

$$\frac{\partial \varepsilon}{\partial \alpha_i} = 2E\left\{\left[x(n) - \sum_{k=1}^{p} \alpha_k\, x(n-k)\right] x(n-i)\right\}$$

$$= 2\left[\sum_{k=1}^{p} E\{\alpha_k\, x(n-k)\, x(n-i)\} - E\{x(n)\, x(n-i)\}\right]$$

$$= 2\left[\sum_{k=1}^{p} \alpha_k\, R(k-i) - R(i)\right]$$

$$= 2 \left\{ \sum_{k=1}^{p} \alpha_k \sum_{n=-\infty}^{\infty} x(n-k)\, x(n-i)\, p[x(n-k), x(n-i)] - \sum_{n=-\infty}^{\infty} x(n)\, x(n-i)\, p[x(n), x(n-i)] \right\}$$

(5.34)

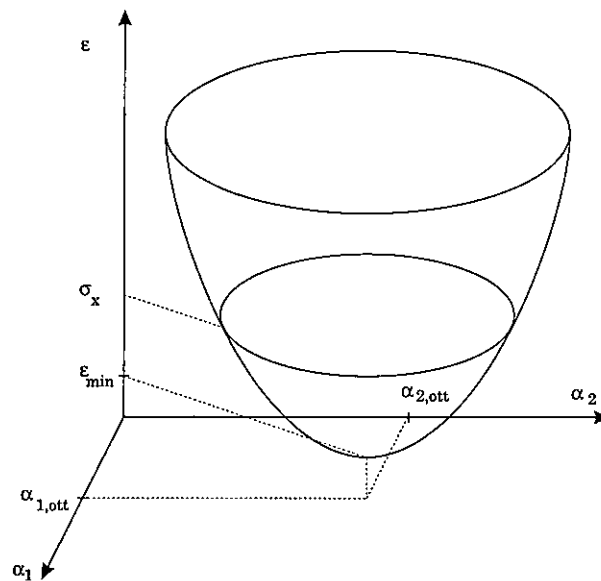



Fig. 5.13 - Superficie dell'errore di predizione per predittore del 2° ordine.

Considerando il sistema di p equazioni ottenute facendo variare l'indice $1 \leq i \leq p$, si ottiene l'espressione del gradiente della funzione d'errore

$$\nabla\varepsilon(\alpha) = 2\,(\mathbf{R}\,\alpha - \mathbf{r}) \tag{5.35}$$

Per ottenerne il minimo è necessario imporre l'annullamento del gradiente. La relazione matriciale ottenuta in tal modo

$$\mathbf{R}\,\alpha = \mathbf{r} \tag{5.36}$$

è detta equazione normale (o di Wiener-Hopf discreta). La soluzione di questo sistema

$$\alpha_{ott} = \mathbf{R}^{-1}\,\mathbf{r} \tag{5.37}$$

fornisce il valore dei coefficienti ottimi del predittore che portano al minimo dell'errore quadratico.
Es.: per il segnale telefonico, i valòri medi normalizzati per l'autocorrelazione a lungo termine possono essere approssimati come R(0) = 1, R(1) = 0.85, R(2) = 0.55, R(3) = 0.25. Per un predittore del terzo ordine, i coefficienti ottimi che ne derivano sono pari a

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & R(2) \\ R(1) & R(0) & R(1) \\ R(2) & R(1) & R(0) \end{bmatrix}^{-1} \begin{bmatrix} R(1) \\ R(2) \\ R(3) \end{bmatrix} = \begin{bmatrix} 1 & 0.85 & 0.55 \\ 0.85 & 1 & 0.85 \\ 0.55 & 0.85 & 1 \end{bmatrix}^{1} \begin{bmatrix} 0.85 \\ 0.55 \\ 0.25 \end{bmatrix} = \begin{bmatrix} 1.45 \\ -0.79 \\ 0.12 \end{bmatrix} \tag{5.38}$$

Sostituendo l'equazione di Wiener-Hopf nell'espressione generale dell'errore, si ricava il suo minimo ε_{min}

$$\varepsilon(\alpha) = \sigma^2 - 2\alpha^T \mathbf{r} + \alpha^T \mathbf{R}\, \alpha = \sigma^2 - \alpha^T \mathbf{r} + \alpha^T (\mathbf{R}\, \alpha - \mathbf{r})$$

$$\varepsilon_{min} = \sigma^2 - \alpha_{ott}^T\, \mathbf{r} = R(0) - \sum_{k=1}^{p} \alpha_k^{ott}\, r(k) \tag{5.39}$$

Esplicitando nell'espressione dell'errore il legame con il suo minimo

$$\varepsilon(\alpha) = \sigma^2 - 2\alpha^T \mathbf{r} + \alpha^T \mathbf{R}\, \alpha = \varepsilon_{min} + \alpha_{ott}^T \mathbf{r} - \alpha^T \mathbf{r} - \mathbf{r}^T \alpha + \alpha^T \mathbf{R}\, \alpha$$

$$= \varepsilon_{min} + \alpha_{ott}^T \mathbf{R} \alpha_{ott} - \alpha^T \mathbf{R} \alpha_{ott} - \alpha_{ott}^T \mathbf{R} \alpha + \alpha^T \mathbf{R}\, \alpha$$

$$= \varepsilon_{min} + (\, \alpha^T \mathbf{R} - \alpha_{ott}^T \mathbf{R}\,)\,(\, \alpha - \alpha_{ott}\,)$$

$$\varepsilon(\alpha) = \varepsilon_{min} + (\, \alpha - \alpha_{ott}\,)^T\, \mathbf{R}\, (\, \alpha - \alpha_{ott}\,) \tag{5.40}$$

Questa relazione può essere semplificata introducendo il vettore di errore dei coefficienti del predittore

$$\mathbf{c} = \alpha - \alpha_{ott} \tag{5.41}$$

La sostituzione di questa variabile nell'espressione della superficie d'errore, equivale a traslare l'origine degli assi in corrispondenza del suo minimo, ottenendo

$$\varepsilon(\alpha) = \varepsilon_{min} + \mathbf{c}^T \mathbf{R}\, \mathbf{c} \tag{5.42}$$

Infine, se si orientano gli assi secondo gli autovettori della matrice **R**, si ottiene la rappresentazione in forma canonica della superficie d'errore. Per far questo è necessario scomporre la matrice di autocorrelazione **R** in funzione della matrice diagonale Λ dei suoi autovalori e della matrice Q dei suoi autovettori

$$\mathbf{R} = \mathbf{Q}\Lambda\mathbf{Q}^T \tag{5.43}$$

Operando il cambiamento di coordinate

$$\mathbf{c} \rightarrow \mathrm{v} = \mathbf{Q}^T \mathbf{c} \tag{5.44}$$

e noto che $\mathbf{Q}^{-1} = \mathbf{Q}^T$, si ottiene, infine, la forma canonica

$$\varepsilon(\alpha) = \varepsilon_{min} + \mathbf{c}^T \mathbf{R}\, \mathbf{c} = \varepsilon_{min} + \mathbf{c}^T \mathbf{Q}\Lambda\mathbf{Q}^T \mathbf{c}$$

$$\varepsilon(\mathrm{v}) = \varepsilon_{min} + \mathrm{v}^T \Lambda\, \mathrm{v} \tag{5.45}$$

Si osserva che la superficie d'errore è legata con legge quadratica allo scostamento dei coefficienti del predittore dagli ottimi tramite la matrice **R**. La convessità del paraboloide è funzione dei suoi autovalori, mentre il loro rapporto ne determina l'eccentricità.

Es.: si consideri la superficie d'errore per un predittore del secondo ordine nel caso di un segnale casuale con σ unitaria.

Gli autovalori della **R** si ottengono annullando il determinante

$$\Delta(\mathbf{R} - \lambda\mathbf{I}) = \Delta\left(\begin{bmatrix} R(0)-\lambda & R(1) \\ R(1) & R(0)-\lambda \end{bmatrix}\right) = [R(0)-\lambda]^2 - R(1)^2 \tag{5.46}$$

ottenendo

$$\lambda_{1,2} = R(0) \pm R(1) \tag{5.47}$$

Per un segnale casuale, la funzione di autocorrelazione è una delta ($R_{xx} = [1\ 0\ 0\ \ldots]$). Di conseguenza gli autovalori risultando coincidenti ed unitari e la superficie risulta essere un paraboloide a sezione circolare (fig. 5.14). Inoltre, per quanto riguarda i coefficienti ottimi del predittore si ottiene

$$\alpha_{ott} = \mathbf{R}^{-1}\mathbf{r} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{5.48}$$

per cui la predizione (ovviamente) non è possibile. Infine, il minimo della funzione d'errore

$$\varepsilon_{min} = \sigma^2 - \alpha_{ott}^T \mathbf{r} = 1 \tag{5.49}$$

coincide con la varianza del segnale. L'errore di predizione, dunque, ha una dinamica maggiore o uguale (nel caso di predizione ottima) a quella del segnale.

Se si considera, invece, un segnale con la stessa varianza, ma con funzione di autocorrelazione $R_{xx} = [\ 1\ 0.75\ 0.25\ \ldots\ ]$, si ottiene

$$\lambda_{1,2} = \begin{bmatrix} 0.25 \\ 1.75 \end{bmatrix}$$

$$\alpha_{ott} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} = \begin{bmatrix} 1.29 \\ -0.71 \end{bmatrix}$$

$$\varepsilon_{min} = 1 - \begin{bmatrix} 1.29 \\ 0.71 \end{bmatrix}^T \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix} = 0.21 \tag{5.50}$$

In tal caso, mantenendo fisso $R(0) = \sigma_x = 1$, la superficie d'errore presenta una sezione ellittica ed il suo minimo risulta inferiore alla varianza del segnale. Nel caso di predizione ottima, quindi, l'errore ha una dinamica inferiore a quella del segnale, per cui la sua codifica può avvenire con un numero inferiore di bit. Se si considera, infine, un segnale con una matrice di autocorrelazione caratterizzata da un rapporto tra autovalori identico a quello appena considerato ($\lambda_1/\lambda_2 = 1/7$), ma con un valore assoluto doppio del precedente

$$\begin{cases} \dfrac{\lambda_1}{\lambda_2} = \dfrac{R(0)-R(1)}{R(0)+R(1)} = \dfrac{1}{7} \\ \lambda_1 = R(0)-R(1) = 2 \times 0.25 \end{cases} \rightarrow \begin{cases} \lambda_1 = 0.5 \\ \lambda_2 = 3.5 \end{cases} \tag{5.51}$$

si ottiene la funzione di auto-correlazione $R_{XX} = [\ 2\ 1.5\ \ldots\ ]$. Imponendo anche la coincidenza sulla posizione del minimo si ottiene

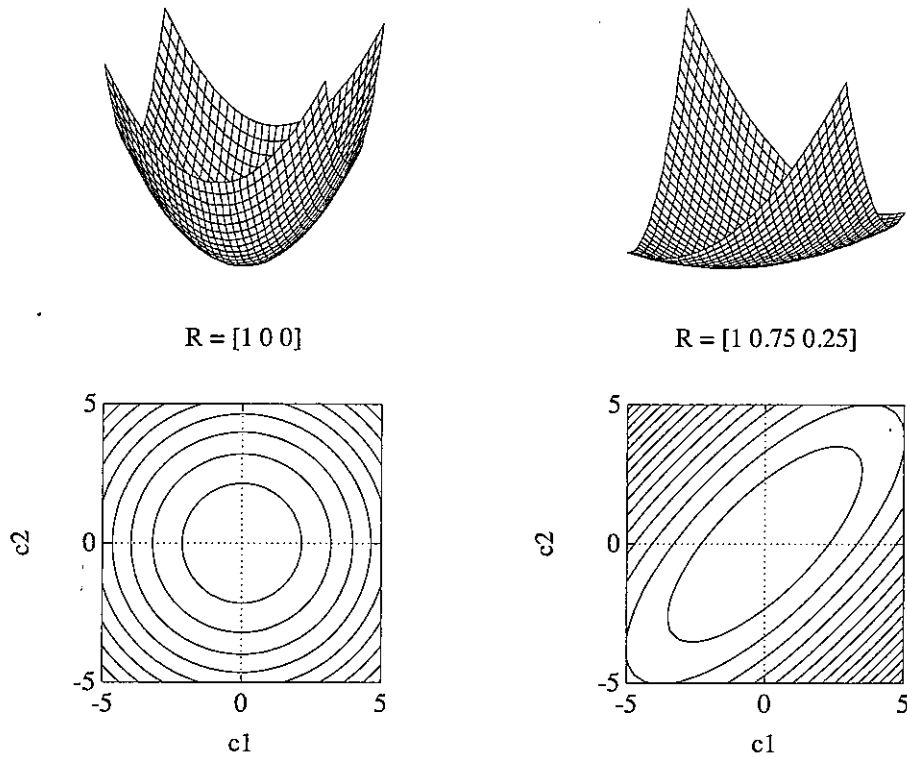



Fig. 5.14 - Superficie del funzionale d'errore.

$$\mathbf{r} = \mathbf{R}\,\alpha_{ott} = \begin{bmatrix} 2 & 1.5 \\ 1.5 & 2 \end{bmatrix} \begin{bmatrix} 1.29 \\ -0.71 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix} \tag{5.52}$$

Con tali valori per la funzione di autocorrelazione (R_{xx} = [2 1.5 0.5 ...]), la superficie d'errore è un paraboloide con la stessa eccentricità di quella ottenuta nel caso precedente (stesso rapporto tra autovalori), ma con una minore concavità (fig. 5.15). Calcolando il suo minimo

$$\varepsilon_{min} = 2 - \begin{bmatrix} 1.29 \\ 0.71 \end{bmatrix}^T \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix} = 0.42 \tag{5.53}$$

si nota come il suo rapporto con la varianza del segnale R(0) risulta identico al caso precedente.

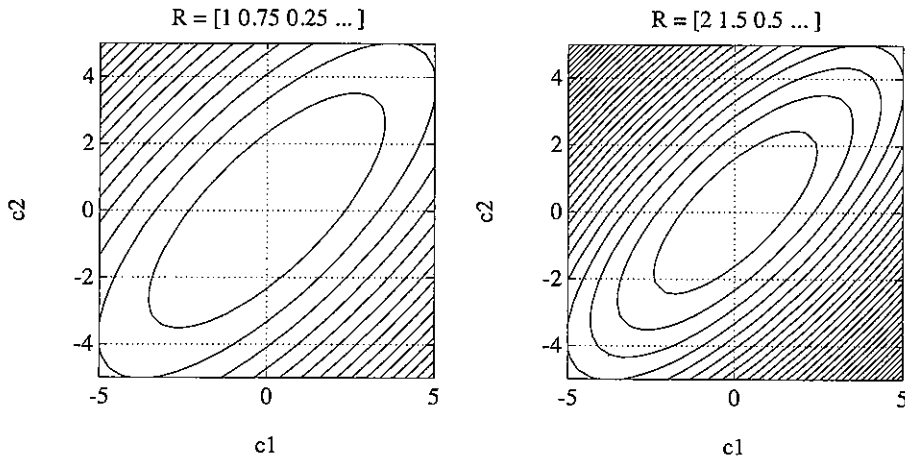



Fig. 5.15 - Superficie del funzionale d'errore.

Volendo analizzare le caratteristiche del segnale d'errore ottenuto nel caso di adattamento ottimo, è opportuno riscrivere l'equazione normale nella forma

$$E\{ \mathbf{x}(n)\, \mathbf{x}^T(n) \}\, \alpha_{ott} = E\{ \mathbf{x}(n)\, \mathbf{x}^T(n-1) \} \tag{5.54}$$

dalla quale si ottiene

$$E\{ \mathbf{x}(n)\, [\, \mathbf{x}^T(n-1) - \mathbf{x}^T(n)\, \alpha_{ott} \,] \} = 0 \tag{5.55}$$

Il termine tra parentesi quadre rappresenta il vettore dei campioni dell'errore di predizione $\mathbf{e}(n)$ (da non confondere con l'errore quadratico ε che è uno scalare), per cui

$$E\{ \mathbf{x}(n)\, \mathbf{e}^T_{ott}(n) \} = 0 \tag{5.56}$$

Tale relazione indica che, nel caso di predizione ottima, l'errore è completamente scorrelato con il segnale (principio di ortogonalità).

Tornando ai legami predittore-modello AR della sorgente, si dimostra che i coefficienti del predittore ottimo coincidono (a meno del segno) con quelli del processo. Infatti, riprendendo l'equazione di Yule-Walker

$$\mathbf{R}_{uu}\, \eta = \mathbf{r}_{uu}\, ; \quad \eta = w \tag{5.57}$$

e riscrivendola per la predizione, in cui u(n) = x(n), ci si riconduce all'equazione di Wiener-Hopf

$$\mathbf{r} = \alpha\,\mathbf{R} \tag{5.58}$$

con la sostituzione $\alpha = -\,\mathbf{w}$. Cioè, nota la struttura della sorgente, il predittore ottimo si ottiene trasformando la struttura del processo sorgente da IIR a FIR e adottando gli stessi pesi, a meno del segno.

Infine, se si confronta l'equazione del processo AR

$$u(n) = v(n) - \sum_{k=1}^{q} w_k\, u(n-k) \tag{5.59}$$

con quella di definizione dell'errore di predizione

$$e(n) = x(n) - \hat{x}(n) \;\;\rightarrow\;\; x(n) = e(n) + \sum_{k=1}^{p} \alpha_k\, x(n-k) \tag{5.60}$$

si nota che, nel caso di predizione ottima, e quindi con $\alpha = -\,\mathbf{w}$, risulta

$$e(n) = v(n) \tag{5.61}$$

cioè l'errore di predizione nel caso di adattamento ottimo coincide con l'ingresso del processo, trasformato da AR a ARX (fig. 5.16). Questo poteva essere intuito pensando di alimentare il processo ARX con un treno di delta. Adottando per il predittore una struttura complementare a quella della sorgente, questo può seguire, una volta che è adattato, la risposta impulsiva della sorgente stessa, ma non può, ovviamente, predire il suo ingresso.

Dato il legame tra i coefficienti del predittore e quelli del processo ARX, l'ordine del predittore non può essere superiore a quello del processo che stima. Altrimenti, la matrice di autocorrelazione R risulta a determinante nullo e l'equazione normale non ammetterebbe soluzione.

Es.: si consideri un segnale sinusoidale del tipo

$$s(n) = \sin\left(n\,\frac{\pi}{4}\right) \tag{5.62}$$

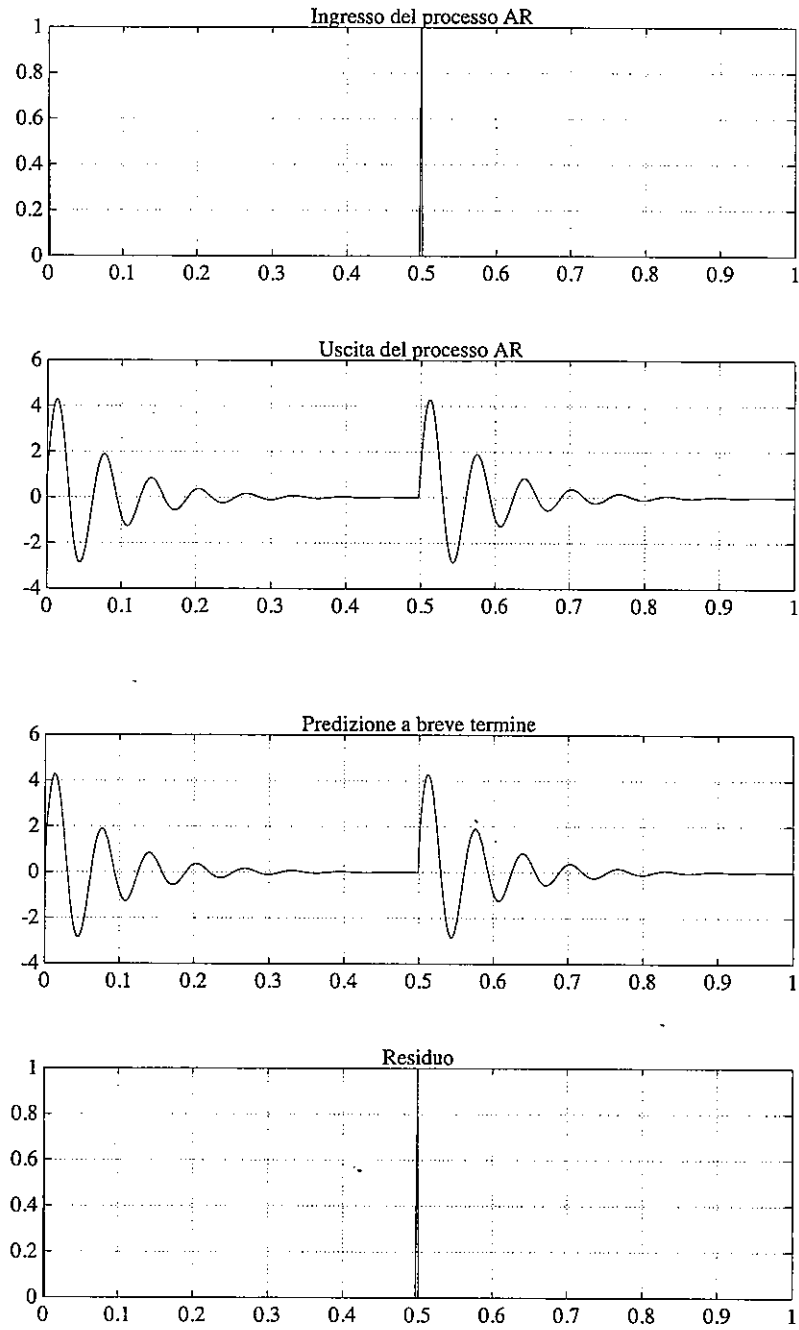



Fig. 5.16 - Legame tra modello ARX e predittore.

Tale segnale può essere pensato come l'uscita di un sistema discreto del secondo ordine con due poli complessi coniugati sulla circonferenza unitaria a $\omega = \pm \pi/4$ [Appendice A.1]. Infatti, data la funzione di autocorrelazione

$$R_{xx}(n) = \frac{1}{2}\cos\left(n\frac{\pi}{4}\right) = \left[\frac{1}{2}, \frac{1}{2\sqrt{2}}, 0, \frac{-1}{2\sqrt{2}}, \frac{-1}{2}, \ldots\right] \quad (5.63)$$

se si considera un predittore del secondo ordine, è possibile ricavarne i relativi coefficienti come

$$\mathbf{r} = \begin{bmatrix} \frac{1}{2\sqrt{2}} \\ 0 \end{bmatrix}; \quad \mathbf{R} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2} \end{bmatrix}; \quad \det(\mathbf{R}) = \frac{1}{8}; \quad \alpha = \begin{bmatrix} \sqrt{2} \\ -1 \end{bmatrix} \quad (5.64)$$

Da tali coefficienti si ottiene un corrispondente processo ARX, il cui denominatore ha radici

$$\mathrm{roots}\left(\begin{bmatrix} 1 \\ \alpha \end{bmatrix}\right) = \begin{bmatrix} \frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} - j\frac{1}{\sqrt{2}} \end{bmatrix} \quad (5.65)$$

che corrisponde a poli con modulo unitario e pulsazione $\omega = \pm \pi/4$. Viceversa, considerando un predittore di ordine maggiore, la matrice di autocorrelazione diviene a determinante nullo e la soluzione dell'equazione di Wiener-Hopf è impossibile. Ad esempio, considerando un predittore del quarto ordine

$$\mathbf{R} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2\sqrt{2}} & 0 & \frac{-1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2} & \frac{1}{2\sqrt{2}} & 0 \\ 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2} & \frac{1}{2\sqrt{2}} \\ \frac{-1}{2\sqrt{2}} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2} \end{bmatrix}; \quad \det(\mathbf{R}) = 0 \quad (5.66)$$

Utilizzare un predittore di ordine eccessivo si traduce in problemi di instabilità numerica nel caso di analisi di segnali con energia concentrata alle frequenze inferiori (e quindi fortemente correlati), in quanto la matrice di

autocorrelazione può divenire mal condizionata [Del93]. Per evitare tali problemi, nel caso di analisi del segnale vocale è prassi comune esaltare le formanti di ordine maggiore applicando una pre-enfasi tramite un filtro passa alto con funzione di trasferimento del tipo

$$P(z) = 1 - \mu z^{-1} \tag{5.67}$$

Una possibile scelta per il parametro μ è

$$\mu = \frac{R_{xx}(1)}{R_{xx}(0)} \tag{5.68}$$

In tal modo l'esaltazione della parte alta dello spettro ha luogo nel caso di componenti vocalizzate, mentre altrimenti risulta trascurabile.

5.2.3 Predizione lineare adattativa a breve termine

Nella codifica predittiva (o quantizzazione differenziale), la stima dei campioni viene sfruttata in trasmissione per ricavare la funzione di errore, che viene codificata e trasmessa al ricevente. Il ricevente ricostruisce il segnale sommando ad una stima eseguita localmente l'errore di predizione ricevuto. Dato che nel caso di predizione ottima su di un segnale correlato, la dinamica dell'errore d(n) è inferiore a quella dell'ingresso x(n), è possibile codificare l'errore con un numero inferiore di bit rispetto a quelli utilizzati per il segnale, riducendo il flusso numerico. Viceversa, mantenendo fisso il numero di bit del quantizzatore, è possibile migliorare il rapporto S/N, che conviene riscrivere come

$$\frac{S}{N} = \frac{E\{x^2(n)\}}{E\{e_q^2(n)\}} = \frac{E\{x^2(n)\}}{E\{d^2(n)\}} \frac{E\{d^2(n)\}}{E\{e_q^2(n)\}} \tag{5.69}$$

dove $e_q(n)$ rappresenta la potenza del rumore di quantizzazione. Il secondo termine del prodotto è il rapporto segnale rumore del quantizzatore, che vede in ingresso il segnale differenza. Il primo termine del prodotto rappresenta il miglioramento del rapporto segnale rumore dovuto alla configurazione differenziale (guadagno di predizione). Questo termine deve essere massimizzato, minimizzando l'errore di predizione.

La trattazione precedente è basata sull'ipotesi di segnale stazionario e funzione di autocorrelazione nota. Se così fosse, stabiliti i coefficienti ottimi del predittore, questi potrebbero rimanere fissi per tutta la codifica del segnale. In pratica i segnali da codificare non sono stazionari, per cui le prestazioni ottenibili sono limitate. Nel caso di segnale vocale, ad esempio, utilizzando coefficienti di predizione fissi derivati dall'andamento medio della funzione di auto-correlazione, non si notano miglioramenti apprezzabili del guadagno all'aumentare dell'ordine del predittore oltre il terzo [Jay84] (fig. 5.17).

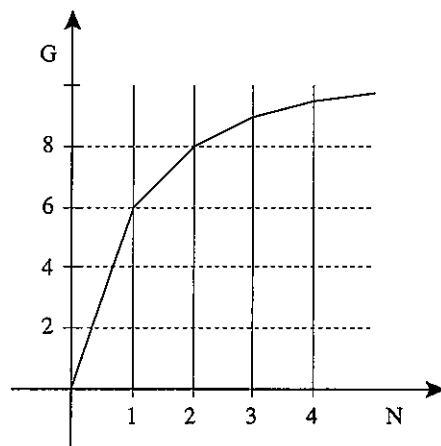



Fig. 5.17 - Guadagno di predizione per predittore a coefficienti fissi.

Per migliorare la predizione è necessario considerare finestre temporali all'interno delle quali il segnale è approssimabile come stazionario e poi aggiornare periodicamente i coefficienti del predittore tramite una stima della funzione di autocorrelazione ricavata dai campioni. Per la stima dell'autocorrelazione (come per la stima del valore efficace nella quantizzazione adattativa) si possono seguire due strade (fig. 5.18). La prima (predizione adattativa in avanti: APF) calcola un'approssimazione della matrice di auto-correlazione a partire da blocchi di campioni del segnale ancora da codificare (predizione a blocchi).

Questo metodo richiede un differente criterio di ottimizzazione per il calcolo dei coefficienti del predittore. In particolare, il criterio scelto è quello di minimizzare non l'errore quadratico medio, ma la sommatoria degli errori quadratici

$$\hat{\varepsilon}(\alpha) = \sum_{n} e(k)^2 = \sum_{n} \left[x(n) - \sum_{k=1}^{p} \alpha_k \, x(n-k) \right]^2 \tag{5.70}$$

dove gli indici che determinano gli estremi della sommatoria in "n", che dipendono dal tipo di algoritmo utilizzato [Appendice C], sono comunque da considerarsi finiti. Tale tecnica è nota come criterio dei minimi quadrati dell'errore (Least Square: LS). Ripetendo passi analoghi a quanto fatto per l'equazione di Wiener-Hopf, si ottiene la seguente equazione normale deterministica

$$\sum_{k=1}^{p} \alpha_k \sum_{n} x(n-i)\, x(n-k) = \sum_{n} x(n-i)\, x(n); \quad 1 \le i \le p \tag{5.71}$$

Tale equazione può essere ottenuta direttamente dall'equazione normale introducendo la grandezza Φ

$$\Phi(i,j) = \frac{1}{n} \sum_{k} x(k-i)\, x(k-j) \tag{5.72}$$

come approssimazione della matrice di autocorrelazione del segnale, ricavata da suoi campioni. Anche in tale equazione gli estremi della sommatoria sono lasciati per il momento indefiniti. L'equazione normale deterministica può essere, quindi, scritto come

$$\sum_{k=1}^{p} \alpha_k \, \Phi(i,k) = \Phi(i,0) \tag{5.73}$$

che, in forma matriciale, diventa

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \Phi(1,3) & \dots & \Phi(1,p) \\ \Phi(2,1) & \Phi(2,2) & \Phi(2,3) & \dots & \Phi(2,p) \\ \Phi(3,1) & \Phi(3,2) & \Phi(3,3) & \dots & \Phi(3,p) \\ \dots & \dots & \dots & \dots & \dots \\ \Phi(p,1) & \Phi(p,2) & \Phi(p,3) & \dots & \Phi(p,p) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \Phi(1,0) \\ \Phi(2,0) \\ \Phi(3,0) \\ \dots \\ \Phi(p,0) \end{bmatrix}.$$

$$\Phi \, \alpha = \psi \tag{5.74}$$

Predizione adattativa in avanti

Predizione adattativa all'indietro

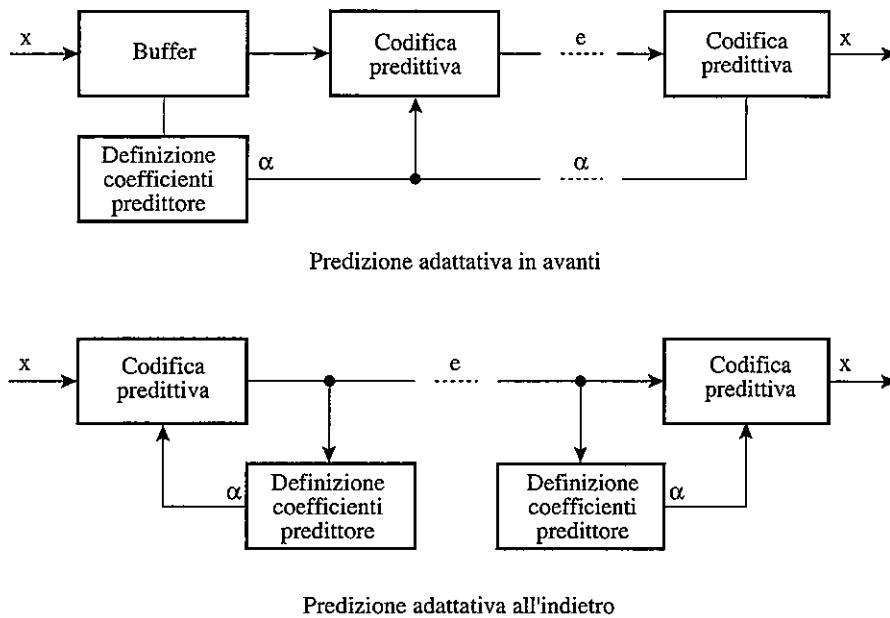



Fig. 5.18 - Predizione adattativa in avanti e all'indietro.

Anche in questo caso, i coefficienti ottimi si ottengono come

$$\alpha_{ott} = \Phi^{1}\, \psi \tag{5.75}$$

L'inversione della matrice di auto-correlazione risulta, in applicazioni real-time, di complessità computazionale eccessiva. Nella predizione a blocchi, quindi, un ruolo importante ha l'individuazione di algoritmi efficienti per la soluzione dell'equazione normale che evitino tale inversione [Appendice C].

Il dover accumulare un blocco di campioni per stimare la funzione di autocorrelazione, comporta un ritardo inevitabile di codifica. Inoltre i coefficienti del predittore vanno trasmessi al ricevente per la decodifica, con le conseguenti ripercussioni sul throughput e sulla robustezza nei confronti di errori di trasmissione. Per tali ragioni, gli algoritmi di predizione a blocchi vengono principalmente utilizzati nella codifica per modelli, dove è essenziale la migliore stima da loro fornita dei parametri del modello della sorgente.

Una seconda famiglia di algoritmi evita di accumulare dati per la stima della funzione di autocorrelazione, richiesta per la soluzione dell'equazione di

Wiener-Hopf. Ciò è possibile utilizzando la finestra degli n campioni che precedono il campione corrente, finestra che viene poi traslata campione per campione (predizione adattativa all'indietro: APB). A partire da una condizione iniziale arbitraria, nella predizione viene utilizzato un vettore sub-ottimo di coefficienti, che sono poi aggiornati ricorsivamente. L'aggiornamento avviene in modo tale che ci si avvicini per approssimazioni successive al minimo dell'errore, muovendoci lungo la sua superficie dell'$\varepsilon(\alpha)$ in direzione opposta a quella del gradiente (algoritmo del gradiente deterministico, o "Steeppest Descent" o "Minimum Mean-Square error gradient algorithm": MMS) (fig. 5.19). In caso di segnale non stazionario, tale minimo risulterà variabile, per cui l'algoritmo continuerà ad inseguirlo continuando ad aggiornare il vettore dei coefficienti.

Ipotizzando, per il momento, nota la matrice di auto-correlazione, il gradiente della funzione di errore

$$\nabla \varepsilon(\alpha) = \frac{\partial \varepsilon}{\partial \alpha_i} = 2(\mathbf{R}\alpha - \mathbf{r}) \qquad (5.76)$$

risulta disponibile. Essendo l'errore una funzione quadratica degli α, il valore del gradiente risulta essere tanto maggiore quanto maggiore è lo scostamento di questi dall'ottimo. Le correzioni da apportare ad essi, quindi, possono essere pensate proporzionali al gradiente stesso tramite una relazione del tipo

$$\alpha^{(n+1)} = \alpha^{(n)} + \frac{1}{2}\mu\,[-\nabla(n)] = \alpha^{(n)} + \mu\left[\mathbf{r} - \mathbf{R}\,\alpha^{(n)}\right] \qquad (5.77)$$

dove μ è opportuno parametro che verrà analizzato successivamente. Il problema è quindi quello del calcolo del gradiente, che è conveniente riscrivere come

$$\nabla\varepsilon(\alpha) = 2(\mathbf{R}\alpha - \mathbf{r}) = 2E\left\{\sum_{i=1}^{p}\left[\sum_{k=1}^{p}\alpha_k\, x(n-k)\, x(n-i) - x(n)\, x(n-i)\right]\right\}$$

$$= -2E\left\{\sum_{i=1}^{p}\left[x(n) - \sum_{k=1}^{p}\alpha_k\, x(n-k)\right]x(n-i)\right\};\quad i = 1, \ldots, p \qquad (5.78)$$

Riconoscendo nel termine tra parentesi quadre l'errore di predizione

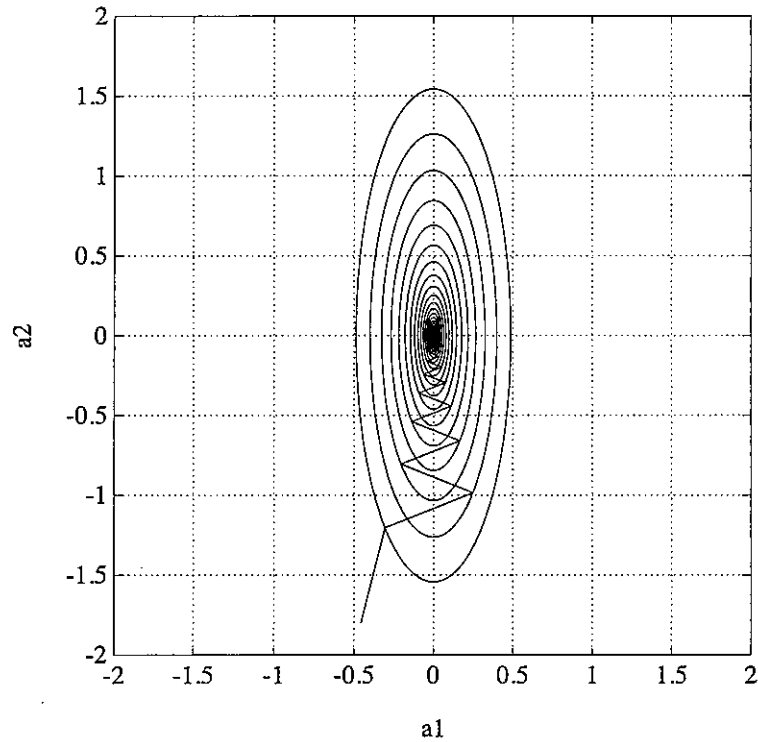



Fig. 5.19 - Ricerca del minimo del funzionale d'errore.

$$e(n) = x(n) - \sum_{k=1}^{p} \alpha_k \, x(n-k) \tag{5.79}$$

si ottiene

$$\nabla \varepsilon(\alpha) \; = \; -2\, E\{\, e(n)\, x(n-i) \,\}\, ; \quad i = 1, ..., p \tag{5.80}$$

Se si definisce $\mathbf{R}_{ex}$ come il vettore di cross-correlazione tra l'errore di stima ed il segnale, questa relazione assume la seguente forma matriciale

$$\nabla \varepsilon(\alpha) = -2\, \mathbf{R}_{ex}(n) \tag{5.81}$$

Questa è la cercata espressione del gradiente del funzionale d'errore. Sostituendola nella relazione ricorsiva per l'aggiornamento dei coefficienti del predittore, si ottiene, infine

$$\alpha^{(n+1)} = \alpha^{(n)} + \mu \, \mathbf{R}_{ex}(n) \tag{5.82}$$

Tale soluzione presuppone che sia nota la matrice $\mathbf{R}_{ex}$. Dato che ciò non si verifica in pratica, sarà necessario sviluppare algoritmi differenti che riescano a stimare il gradiente dai campioni del segnale disponibili, come mostrato nel successivo sottoparagrafo.

Nella legge di aggiornamento, il parametro μ influenza la memoria dell'algoritmo, dato che esso può essere interpretato come coefficiente di feed-back per il vettore dei coefficienti e quindi come coefficiente di smorzamento sul contributo dei precedenti campioni. La scelta parametro μ, quindi, è estremamente critica. Scegliendo valori "piccoli" si ha una convergenza lenta dell'algoritmo, a causa delle modeste correzioni apportate al vettore dei coefficienti ad ogni passo dell'algoritmo. Viceversa, valori "grandi" di μ, rendono l'algoritmo instabile, dato che si continua a correggere energicamente il vettore dei coefficienti anche quando si è in prossimità del minimo.

Per studiare la stabilità dell'algoritmo è necessario ricavare una relazione ricorsiva per il vettore di errore dei coefficienti del predittore, il cui termine al passo n-esimo è definito come

$$\mathbf{c}^{(n)} = \alpha^{(n)} - \alpha_{ott} \tag{5.83}$$

Il vettore di errore al passo n+1 si ottiene come

$$\mathbf{c}^{(n+1)} = \alpha^{(n+1)} - \alpha_{ott} = \alpha^{(n)} - \mu \, (\mathbf{R} \, \alpha^{(n)} - \mathbf{r} \,) - \alpha_{ott}$$

$$= \alpha^{(n)} - \alpha_{ott} - \mu \, \mathbf{R} \, \alpha^{(n)} + \mu \, \mathbf{r} = \mathbf{c}(n) - \mu \, \mathbf{R} \, \alpha^{(n)} + \mu \, \mathbf{R} \, \alpha_{ott}$$

$$\mathbf{c}^{(n+1)} = (\mathbf{I} - \mu \, \mathbf{R}) \, \mathbf{c}^{(n)} \tag{5.84}$$

Scomponendo la R in funzione dei suoi autovalori

$$\mathbf{c}^{(n+1)} = (\mathbf{I} - \mu \, \mathbf{Q} \, \Lambda \, \mathbf{Q}^T) \, \mathbf{c}^{(n)}$$

$$\mathbf{Q}^T \, c(n+1) = (\mathbf{I} - \mu \, \Lambda \,) \, \mathbf{Q}^T \, c(n) \tag{5.85}$$

e trasformando la superficie del criterio d'errore in forma canonica, tramite il cambiamento di coordinate

$$c^{(n)} \rightarrow \nu^{(n)} = Q^T c^{(n)} = Q^T \left[\alpha^{(n)} - \alpha_{ott} \right] \tag{5.86}$$

si ottiene la seguente equazione ricorsiva per il vettore ν

$$Q\, \nu^{(n+1)} = (I - \mu\, R)\, Q\nu^{(n)}$$

$$\nu^{(n+1)} = (\, I - \mu\, \Lambda\,)\, \nu^{(n)} \tag{5.87}$$

Esprimendo scalarmente tale relazione ricorsiva, si ottiene

$$\nu_k^{(n+1)} = (\,1 - \mu\, \lambda_k\,)\, \nu_k^{(n)}$$
$$\nu_k^{(n)} = (\,1 - \mu\, \lambda_k\,)^n\, \nu_k^{(0)} \tag{5.88}$$

Questa rappresenta una serie geometrica, la cui stabilità è assicurata se risulta

$$-1 < 1 - \mu\, \lambda_k < 1$$
$$0 < \mu < \frac{2}{\lambda_{max}} \tag{5.89}$$

Questa relazione, che impone un limite superiore al valore della costante di aggiornamento, mostra come la μ debba essere scelta in funzione degli autovalori della matrice di autocorrelazione del processo, peraltro incogniti. Sostituendo l'espressione ricorsiva della variabile ν in quella dell'errore, si ricava infine

$$\varepsilon(\nu) = \varepsilon_{min} + \nu^T \Lambda\, \nu = \varepsilon_{min} + \sum_{k=1}^{p} \lambda_k\, \nu_k^2 = \varepsilon_{min} + \sum_{k=1}^{p} (1 - \mu\, \lambda_k)^{2n}\, \nu_k(0)^2 \tag{5.90}$$

che mostra come l'errore tenda al suo minimo con legge esponenziale.

Per quanto riguarda la rapidità di convergenza, la dipendenza dagli autovalori della matrice di autocorrelazione è evidente, dato che da essi dipende l'eccentricità delle sezioni orizzontali del funzionale d'errore. È intuitivo pensare, infatti, che con sezioni approssimativamente circolari in gradiente punti al minimo del funzionale d'errore e quindi le correzioni dell'algoritmo portano i coefficienti del predittore a raggiungerlo rapidamente. Con sezioni fortemente ellittiche, invece, le correzioni apportate ai parametri hanno forti componenti trasversali alla direzione di massima pendenza, per cui l'avvicinamento al minimo sarà più lento.

5.2.4 Metodo del gradiente

Con l'algoritmo del gradiente deterministico, il predittore utilizza un vettore di coefficienti che viene periodicamente aggiornato. A tal fine ci si muove lungo la superficie del criterio d'errore verso il suo minimo, andando nella direzione opposta al gradiente

$$\nabla\varepsilon(\alpha) = \frac{\partial\varepsilon}{\partial\alpha_i} = 2\,(\mathbf{R}\,\alpha - \mathbf{r}) \tag{5.91}$$

Ciò si ottiene tramite una legge di aggiornamento del tipo

$$\alpha^{(n+1)} = \alpha^{(n)} + \frac{1}{2}\,\mu\,[\,-\nabla(n)\,] = \alpha^{(n)} + \mu\left[\,\mathbf{r} - \mathbf{R}\,\alpha^{(n)}\,\right] \tag{5.92}$$

con μ un opportuno parametro. Ipotizzando nota la matrice di autocorrelazione, il gradiente della funzione di errore risulta disponibile e calcolabile in funzione dei coefficienti del predittore α_i tramite il valore atteso

$$\nabla\varepsilon(\alpha) \;=\; -2\,E\{\,e(n)\,x(n-i)\,\}\;;\quad i = 1, ..., p \tag{5.93}$$

Il metodo del gradiente stocastico, detto brevemente metodo del gradiente (Least Mean Square error gradient algorithm: LMS), elimina la necessità di disporre di tale valore atteso, ricorrendo ad una sua stima istantanea, ottenuta dai campioni del segnale come

$$\nabla\hat{\varepsilon}(\alpha) = -\,2\,e(n)\,x(n-k) \tag{5.94}$$

Questa semplificazione porta a delle relazioni ricorsive per il calcolo dei coefficienti del predittore del tipo

$$\alpha_k^{(n+1)} = \alpha_k^{(n)} + \mu\,e(n)\,x(n-k) \tag{5.95}$$

dove l'errore al passo n-esimo si ottiene come differenza tra l'ingresso e la sua stima

$$e(n) = x(n) - \sum_{k=1}^{p} \alpha_k^{(n)}\,x(n-k) \tag{5.96}$$

Es.: si consideri un segnale cosinusoidale campionato con una frequenza pari a quattro volte la frequenza del segnale e fase iniziale nulla. La sequenza dei campioni è pari a

$$x(n) = [\ 1\ 0\ -1\ 0\ 1\ 0\ -1\ \ldots\ ];\quad n = 0, 1, \ldots \tag{5.97}$$

Si consideri un predittore del 2° ordine con costante di aggiornamento dei coefficienti $\mu = 1/2$ e condizioni iniziali per i coefficienti del predittore $\alpha = [\ 0\ 0\ ]$. Per i primi due passi dell'algoritmo si ottiene

$$\begin{gathered} e(0) = x(0) = 1;\ e(1) = x(1) = 0 \\ \alpha_1^{(1)} = \mu\ x(0)\ e(1) = 0 \end{gathered} \tag{5.98}$$

dopo di che si possono applicare le relazioni ricorsive esposte che, per n = 2, forniscono

$$e(2) = x(2) - \alpha_1^{(2)}\ x(1) = -1$$

$$\alpha_1^{(3)} = \alpha_1^{(2)} + \mu\ e(2)\ x(1) = 0;\ \alpha_2^{(3)} = \alpha_2^{(2)} + \mu\ e(2)\ x(0) = -\frac{1}{2} \tag{5.99}$$

per n = 3 forniscono

$$e(3) = x(3) - \alpha_1^{(3)}\ x(2) - \alpha_2^{(3)}\ x(1) = 0$$

$$\alpha_1^{(4)} = \alpha_1^{(3)} + \mu\ e(3)\ x(2) = 0;\ \alpha_2^{(4)} = \alpha_2^{(4)} + \mu\ e(3)\ x(1) = -\frac{1}{2} \tag{5.100}$$

e cosi via.

Per quanto riguarda la convergenza dell'LMS, essa, come per l'MMS, è di tipo esponenziale con decadimento regolato da μ (fig. 5.20). L'uso di una stima del gradiente invece del suo valore corretto, però, si traduce nell'aggiunta di un errore che si manifesta come fluttuazioni sulla convergenza esponenziale. Di conseguenza, da μ dipende anche lo scostamento dal valore teorico del minimo errore ottenibile. Infatti, con valori bassi di μ si ottiene una convergenza lenta, ma i valori dei coefficienti ottenuti risultano essere molto vicini a quelli ottimi; con valori di μ elevati la convergenza è più rapida, ma le maggiori fluttuazioni introdotte fanno si che le prestazioni a regime siano peggiori. La soluzione può essere quella di rendere adattativo anche tale parametro, al prezzo di una maggiore complessità computazionale [Appendice B].

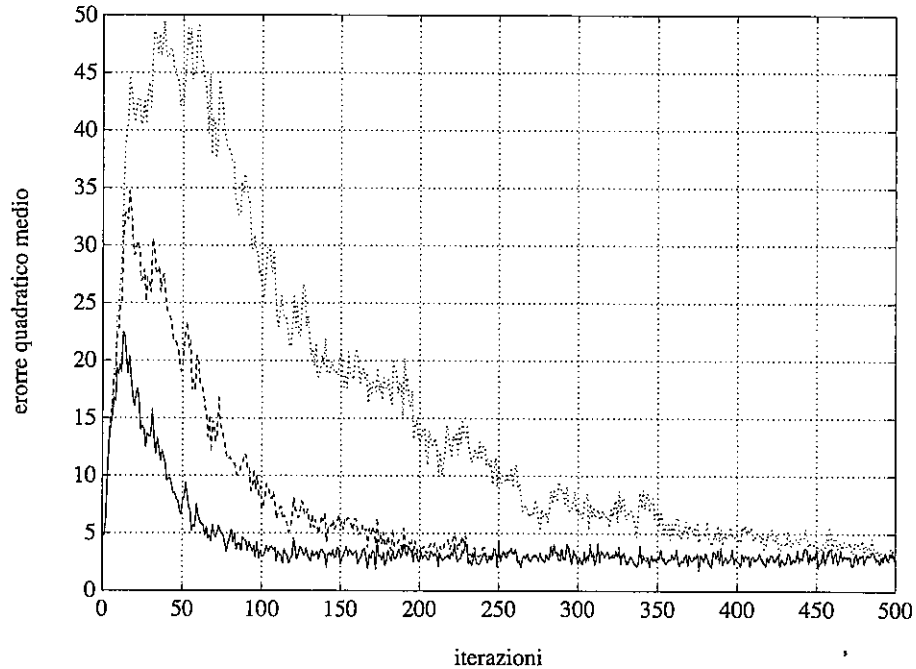



Fig. 5.20 - Curve di apprendimento.

Al fine di ridurre la complessità di implementazioni hardware dell'algoritmo, è possibile adottare leggi di aggiornamento dei coefficienti del predittore che si basino esclusivamente sul segno della funzione d'errore e/o del segnale (polarity cross-correlations). Le leggi di aggiornamento che ne derivano sono del tipo

$$\alpha_k^{(n+1)} = \alpha_k^{(n)} + \mu \operatorname{sgn}[\, e(n) \,]\, x(n-k)$$

$$\alpha_k^{(n+1)} = \alpha_k^{(n)} + \mu\, e(n) \operatorname{sgn}[\, x(n-k) \,]$$

$$\alpha_k^{(n+1)} = \alpha_k^{(n)} + \mu \operatorname{sgn}[\, e(n) \,] \operatorname{sgn}[\, x(n-k) \,] \tag{5.101}$$

Infine, per ridurre la propagazione di errori di quantizzazione o di trasmissione, può risultare utile utilizzare un coefficiente moltiplicativo $\lambda < 1$ per attenuare il peso dei precedenti coefficienti α (potenzialmente errati), ottenendo

$$\alpha_k^{(n+1)} = \lambda\ \alpha_k^{(n)} + \mu\ \text{sgn}[\ e(n)\ ]\ \text{sgn}[\ x(n-k)\ ] \tag{5.102}$$

In tal modo, l'influenza di un errore sul valore di un coefficiente viene via via attenuato man mano che la stima prosegue.

5.2.5 Coefficienti di predizione lineare a breve termine

Nei paragrafi precedenti si è visto che le metodologie di analisi spettrale e gli strumenti di calcolo associati (descritti in Appendice C) abbiano avuto un ruolo molto importante nel settore della codifica del segnale vocale. Più in generale si può affermare che l'analisi spettrale a breve termine è stato, ed è ancora, uno strumento essenziale per lo studio del segnale vocale ed è stato impiegato in tutti i settori del trattamento del segnale vocale, comprendendo quindi non solo la codifica, ma anche il riconoscimento e la sintesi.

I passi più significativi del successo di questa tecnologia sono riassunti in tabella 5.2.

EVENTO	ANNO
Koenig inventa lo spettrografo (Sound Spectrograph)	1946
Fant pubblica la teoria acustica della produzione della voce	1960
In M.I.T. ed ai laboratori Bell si introduce il metodo di analisi per sintesi per il calcolo del modello spettrale e delle formanti	1961 - 1964
Makhoul,, Markel,, ed altri propongono il modello di analisi con predizione lineare (LPC - Linear Predictive Coding)	1968 - 1975

Tab. 5.2 - Calendario degli eventi legati alle tecniche di analisi spettrale.

Si può senz'altro affermare che la svolta per l'utilizzo generalizzato dell'analisi spettrale avviene negli anni settanta con l'avvento della tecnica LPC (Linear Predictive Coding), la quale consente due grandi vantaggi rispetto ai metodi proposti precedentemente:

- il calcolo dei parametri è diretto e non iterativo, come era nei metodi di analisi per sintesi;
- il calcolo dei parametri è fatto usando la rappresentazione del segnale nel tempo.

Anche in questo caso, la disponibilità di uno strumento di calcolo semplice ha stimolato la ricerca in molti settori del signal processing e non ultimo in quello della codifica. In particolare, l'analisi spettrale LP è stata ampiamente utilizzata per sfruttare una caratteristica fondamentale del segnale vocale e cioè quella di avere uno spettro di ampiezza non uniforme in frequenza ed in prima approssimazione stazionario a breve termine. Questa caratteristica è stata sfruttata in due modi distinti in due filoni specifici della codifica:

- Codificatori predittivi - In questo caso si osserva che le caratteristiche spettrali del segnale vocale si traducono in una correlazione nel tempo tra campioni adiacenti. Ne consegue la possibilità di effettuare una predizione adattativa del segnale con conseguente riduzione di ridondanza.
- Codificatori per modelli - In questo caso le tecniche di codifica sono basate sull'impiego di un modello a parametri adattativi e le caratteristiche spettrali del segnale vocale sono legate ai parametri del tubo acustico nel modello di produzione del segnale vocale.

Le tecniche di predizione lineare adattativa viste in precedenza sfruttano la correlazione nel tempo a breve termine tipica del segnale vocale. Tale correlazione viene sfruttata calcolando un segnale predetto che viene sottratto al segnale originale e trasmettendo quindi il solo segnale errore di predizione. Come si vedrà nel caso della tecnica di codifica DPCM, il segnale predetto, e cioè la stima del segnale originale, è calcolato a partire dal segnale ricostruito, al fine di garantire la completa reversibilità della codifica predittiva (stima in catena chiusa). Ne consegue che in questo caso i coefficienti del filtro di predizione sono calcolabili sia al trasmettitore che al ricevitore e quindi non devono essere trasmessi.

Viceversa nel caso di schemi di codifica per modelli i coefficienti del filtro a soli-poli, relativi al modello del tratto vocale (coincidenti con i coefficienti di predizione quando il metodo di calcolo è basato sulla minimizzazione dell'errore di predizione), sono calcolati a partire dal segnale originale e sono quindi trasmessi al ricevitore dove sono utilizzati nel filtro di sintesi. È evidente che in questo caso tali coefficienti devono essere opportunamente codificati e trasmessi come informazione ausiliaria, necessaria al fine di sintetizzare il segnale vocale.

Vista la necessità di trasmettere tali coefficienti, sono state ipotizzate diverse trasformazioni degli stessi, al fine di ottenere un insieme di coefficienti che abbia caratteristiche ideali dal punto di vista della sensibilità al rumore di quantizzazione, delle proprietà di interpolazione degli spettri e della stabilità del filtro corrispondente.

Nel seguito sono riportate le trasformazioni di coefficienti più utilizzate nel campo della codifica, mettendo in luce per ognuna di esse quali sono le caratteristiche peculiari.

Coefficienti di autocorrelazione

Sono i coefficienti di autocorrelazione del segnale di riferimento e sono quindi calcolati con la relazione

$$R_k = \sum_{n=0}^{L_H-k} x(n) \cdot x(n+k) \qquad (5.103)$$

in cui x(n) rappresenta l'n-esimo campione del segnale di riferimento ed L_H la lunghezza della finestra di analisi. Non hanno caratteristiche di rilievo, eccettuata le caratteristiche di interpolazione. Solitamente costituiscono il primo passo per calcolare altri insiemi di coefficienti.

Coefficienti del filtro diretto

Costituiscono i coefficienti del filtro diretto ai. Possono essere calcolati agevolmente dai coefficienti di autocorrelazione risolvendo la matrice di Yule-Walker [Appendice C.1] oppure con l'algoritmo recursivo di Levinson-Durbin [Appendice C.2]. Hanno caratteristiche di robustezza al rumore di quantizzazione abbastanza scadenti e quindi non sono stati quasi mai utilizzati direttamente per la trasmissione. Viceversa hanno il pregio di poter essere impiegati direttamente in una struttura di filtro numerico che è quella del filtro diretto. Hanno delle buone caratteristiche di interpolazione ma non esiste una formula diretta per valutare se un set di coefficienti corrisponde ad un filtro stabile.

Coefficienti di riflessione (PARCOR)

Possono essere ottenuti direttamente dai coefficienti α_i tramite l'algoritmo di Levinson-Durbin, [Appendice C.3] oppure utilizzando la recursione di Schur [Appendice C.4]. Devono il loro nome di PARCOR (PARtial CORrelation) al metodo di calcolo. Anche per questi coefficienti esiste una struttura di filtro implementabile che è quella a traliccio. Hanno una bassa sensitività al rumore di quantizzazione, in confronto ai coefficienti del filtro diretto, ed hanno la proprietà di avere una condizione di stabilità del filtro data da

$$\text{stabilità} \Leftrightarrow |k_i| < 1 \supset i \qquad (5.104)$$

Per contro le proprietà di interpolazione di spettri non sono molto buone.

Area functions

Sono coefficienti legati ai coefficienti di riflessione tramite la relazione

$$A_i = A_{i+1} \cdot \frac{1+k_i}{1+k_i} \text{ con } i = p, p-1, \dots, 1 \text{ e } A_{p+1} = 1 \qquad (5.105)$$

dove p rappresenta l'ordine di predizione del filtro. Hanno il significato fisico di essere legati all'area delle sezioni trasversali del tubo acustico (nel modello del tratto vocale) da cui hanno tratto il nome. Hanno discrete proprietà di sensitività agli errori ma per contro pessime proprietà di interpolazione.

Log Area Ratios (LAR)

Questi coefficienti sono stati utilizzati molto frequentemente in passato a causa delle buone proprietà di sensitività agli errori di quantizzazione. In particolare la sensitività agli errori è circa uguale per i diversi coefficienti. Le prestazioni in termini di interpolazione sono simili a quelle dei coefficienti di riflessione.

Sono ricavati dai coefficienti di riflessione tramite la relazione

$$LAR_i = \ln\left[\frac{A_{i+1}}{A_i}\right] = \ln\left[\frac{1-k_i}{1-k_i}\right] \quad 1 \le i \le p \qquad (5.106)$$

ed hanno quindi il significato del rapporto logaritmico tra due aree adiacenti del tubo acustico senza perdite che è usato per modellare il tratto vocale. Anche in questo caso esiste una condizione di stabilità che è data da

$$\text{stabilità} \Leftrightarrow \mathrm{LAR}_i > 0 \tag{5.107}$$

È immediato verificare che tale condizione di stabilità discende direttamente da quella sui coefficienti di riflessione.

Modified Log Area Ratios (MLAR)

Sono molto simili ai LAR ed hanno trovato scarsissimo utilizzo. Sono stati introdotti in letteratura al fine di avere una maggiore risoluzione dei primi due coefficienti, in considerazione del fatto che solitamente tali coefficienti sono prossimi a 1. Si calcolano con la relazione

$$\mathrm{MLAR}_i = \ln\left[\frac{F-k_i}{F-k_i}\right] \quad 1 \le i \le p \;\; \text{e} \;\; F > 1 \tag{5.108}$$

Inverse Sin Coding (ISC)

Anche questi coefficienti sono stati impiegati abbastanza saltuariamente. Sono definiti dalla relazione

$$\mathrm{ISC}_i = \arcsin(k_i) \tag{5.109}$$

Coefficienti cepstrali

Sono anche chiamati LPC cepstrum e sono i coefficienti cepstrali dell'inviluppo spettrale derivato dall'analisi LPC. Pertanto non sono uguali ai coefficienti cepstrali dello spettro del segnale vocale. In pratica si possono calcolare con la relazione recursiva

$$c_i = \alpha_i + \sum_{j=1}^{i-1} \frac{j}{i} c_j \, \alpha_{i-j} \quad 1 \le i \le p \;\; \text{e} \;\; \alpha_0 = 1 \tag{5.110}$$

a partire dai coefficienti del filtro diretto α_i. Il coefficiente c_0 solitamente è posto pari al logaritmo del guadagno del filtro di sintesi. Le prestazioni non sono eccezionali e pertanto sono stati utilizzati poco nella codifica, mentre hanno trovato largo impiego in altri settori, come il riconoscimento.

Line Spectrum Pairs (LSP)

Una menzione particolare spetta ai coefficienti Line Spectrum Pair o coefficienti LSP in quanto sono stati impiegati in molti schemi di codifica, in considerazione delle loro buone prestazioni.

La loro introduzione si deve principalmente a Wakita [Wak81] (anche se la paternità del concetto spetta ad Itakura nel 1975 [Ita75]) mentre Kabal e Ramachandran hanno pubblicato un algoritmo efficiente di calcolo basato sui polinomi di Chebyshev [Kab86].

I coefficienti LSP possono essere presentati in modo agevole partendo dalla solita equazione del filtro di sintesi che definisce un modello spettrale a soli poli di ordine p

$$A(z) = 1 - \sum_{i=1}^{p} \alpha_i z^{-i} \tag{5.111}$$

A partire da questa equazione si possono definire due polinomi di ordine (p+1), uno simmetrico P(z) ed uno antisimmetrico Q(z), in base alle relazioni

$$P(z) = A(z) - z^{-(p+1)} \cdot A(z^{-1})$$

$$Q(z) = A(z) + z^{-(p+1)} \cdot A(z^{-1}) \tag{5.112}$$

Dalle due relazioni è facile verificare che

$$A(z) = \frac{1}{2} [P(z) + Q(z)] \tag{5.113}$$

Le radici dei due polinomi ausiliari P(z) e Q(z) si chiamano Line Spectrum Pairs (LSP), mentre le posizioni angolari, sul piano complesso, di tali radici si chiamano Line Spectrum Frequencies (LSF).

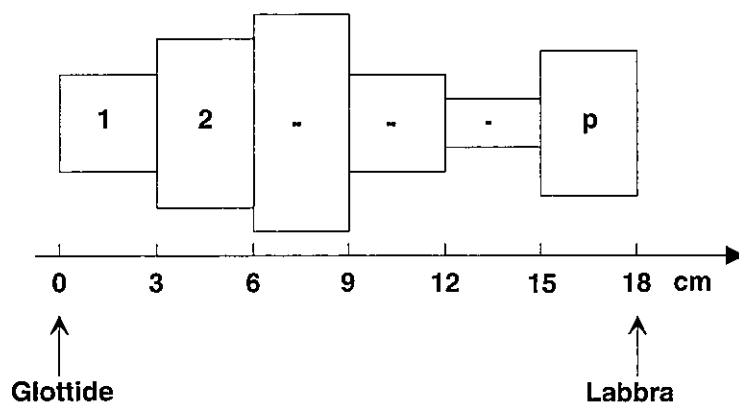



Fig. 5.21 - Rappresentazione grafica del modello di tubo acustico.

I coefficienti LSP (o coppie di linee spettrali) hanno un significato fisico ben preciso che può essere enunciato facendo riferimento al modello acustico del tratto vocale. Tale modello, nella sua rappresentazione più semplice, è costituito da un tubo acustico caratterizzato, nel caso di ordine di predizione p, da p sezioni equispaziate (fig. 5.21).

Il tubo acustico ha lunghezza pari alla lunghezza del tratto vocale e quindi si sviluppa dalla glottide alle labbra. Le p sezioni hanno area opportuna in considerazione dei processi articolatori e quindi dei suoni emessi. Inoltre il tubo acustico è adattato alla glottide, dove viene alimentata l'eccitazione, e si apre su una sezione infinita in corrispondenza delle labbra.

I coefficienti di riflessione (k_i), prima introdotti, tengono conto del rapporto di energie tra onda riflessa ed onda incidente alle varie sezioni, o, analogamente, del disadattamento di impedenza nel passare da una sezione a quella adiacente. In questa esemplificazione i due termini prima introdotti alla funzione A(z) per ottenere P(z) e Q(z), corrispondono a considerare uno stadio aggiuntivo, in corrispondenza della glottide, con $k_{p+1} = \pm 1$ e cioè uno stadio completamente aperto o completamente chiuso.

In queste condizioni il tubo acustico è senza perdite, quindi il fattore di risonanza diventa infinito e lo spettro della funzione di trasferimento risulta costituito, in questo caso, da un insieme di delta di Dirac. In tal caso quindi, le LSP, e cioè le radici di questa funzione di trasferimento, assumono il significato di *coppie di linee spettrali*.

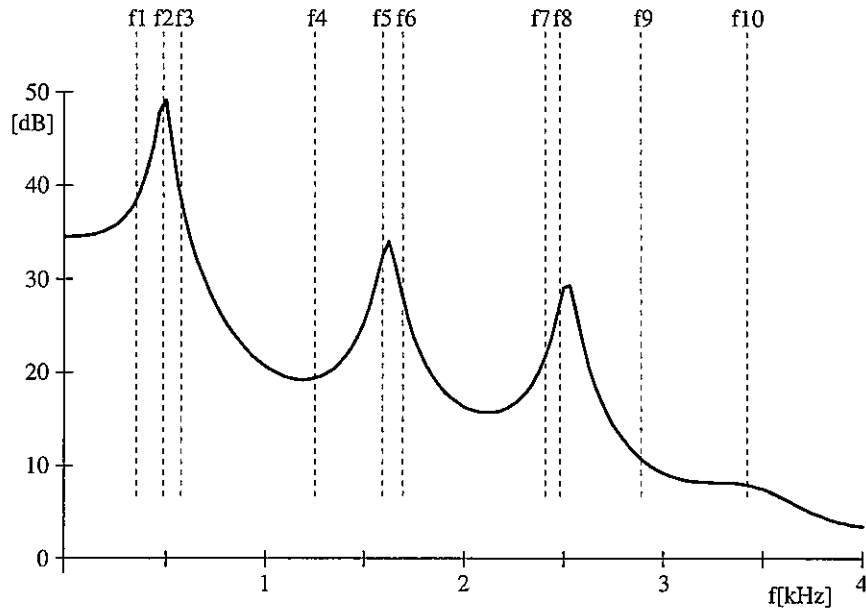



Fig. 5.22 - Inviluppo spettrale LPC di ordine 10 e relative LSF per un tratto di voce vocalizzata.

Un tipico esempio di spettro di sintesi LPC e frequenze di linee spettrali (LSF) è riportato in figura 5.22. Da questa figura, ed in considerazione del significato fisico illustrato, si evince una stretta relazione tra i coefficienti LSP e la posizione delle formanti dello spettro. Questa relazione è alla base della migliore efficienza che si ottiene con le LSP nel codificare l'informazione spettrale.

La relazione tra LSP e formanti può essere espressa con le seguenti considerazioni qualitative.

Le LSP tendono a raggrupparsi nell'intorno delle formanti.

Più il fattore di risonanza (Q) è alto, più le LSP sono vicine.

Soong e Juang in [Soo84] hanno dimostrato che se A(z) è a fase minima (e cioè relativa ad un filtro stabile) allora le radici di P(z) e Q(z) giacciono sul cerchio unitario, sono complesse coniugate e sono alternate (cioè soddisfano la condizione $0 \leq \omega_{q0} \leq \omega_{p0} \leq \omega_{q1} \leq \omega_{p1} \leq \ldots \leq 2\pi$). Come conseguenza diretta, ogni insieme di radici alternate, limitate nell'intervallo [0, 2π]), è relativo ad un filtro stabile. Quest'ultima considerazione fornisce quindi un criterio di stabilità per questo insieme di coefficienti.

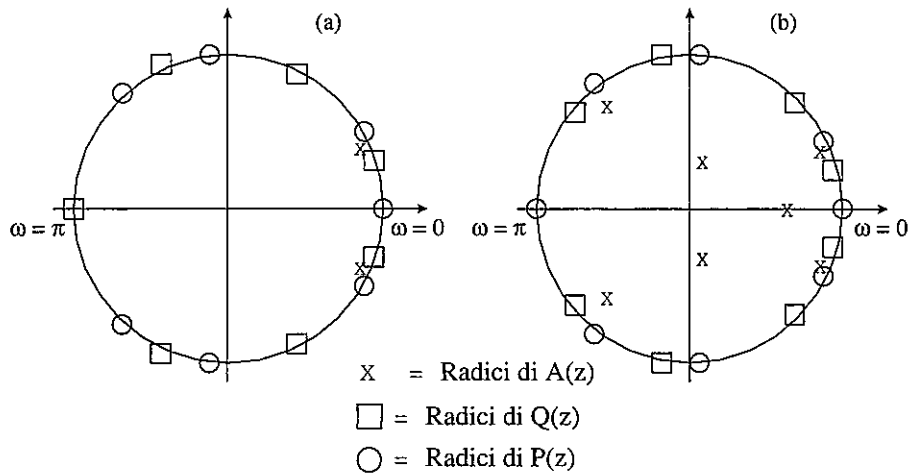



Fig. 5.23 - Luogo delle radici e rappresentazione delle LSP. (a) ordine 6 e (b) ordine 7.

A titolo esemplificativo la figura 5.23 riporta il luogo delle radici per due generici set di LSP nel caso di ordine pari (5.23a) ed ordine dispari (5.23b).

Nella figura sono anche riportate le radici del filtro di sintesi. Si può facilmente osservare che per radici di A(z) prossime al cerchio unitario le LSP tendono ad avvicinarsi tra loro. Viceversa, la vicinanza tra coppie di LSP non è necessariamente un buon indicatore della vicinanza di una radice di A(z) al cerchio unitario e cioè di una risonanza.

I metodi di calcolo delle radici LSP sfruttano le caratteristiche enunciate precedentemente. Dato che le LSP occorrono in coppie complesse coniugate, è sufficiente calcolare le radici su di un semicerchio, ad esempio nell'intervallo $0 \div \pi$. Questo rimuove l'apparente aumento di ridondanza nel passare da A(z) ai due polinomi P(z) e Q(z). Le radici possono essere ricavate direttamente a partire dalle equazioni con i metodi dell'aritmetica complessa, oppure in modo molto più agevole considerando che le radici giacciono sul semicerchio unitario e quindi ponendo $z = e^{j\omega}$ e risolvendo per ω. Questo metodo richiede tuttavia lo sviluppo e la soluzione di espressioni trigonometriche anche complesse. Un metodo di calcolo più semplice che si presta ad essere svolto numericamente è quello basato sull'utilizzo dei polinomi di Chebyshev [Kab86]. In questo caso i termini trigonometrici $\cos(m\omega)$ sono sostituiti con i polinomi $Tm(x)$, sfruttando la mappatura $x = \cos(\omega)$. Ne consegue che le radici nel dominio

di x sono confinate nell'intervallo [+1,-1]. Con tali posizioni, il metodo di calcolo consiste nel determinare, per opportuni incrementi, il cambiamento di segno del polinomio e quindi rifinire la ricerca con bisezioni successive.

Interpolazione di spettri

I parametri LPC, comunque trasformati, sono impiegati in codificatori cosiddetti parametrici. In considerazione della stazionarietà a breve termine del segnale vocale, tali coefficienti sono da considerarsi rappresentativi delle caratteristiche spettrali del segnale vocale per un intervallo di tempo limitato.

Questo intervallo (frame) è in generale di durata compresa tra 2 e 40 ms. Dato che i parametri LPC devono essere aggiornati e trasmessi ad ogni frame, un frame lungo è spesso richiesto per consentire una bassa velocità di trasmissione. Tuttavia un frame troppo lungo determina parametri spettrali non sufficientemente rappresentativi della dinamica articolatoria, con conseguente introduzione di distorsioni spettrali. Inoltre la lunghezza del frame di analisi determina un contributo fisso al ritardo algoritmico della tecnica di codifica ed anche il ritardo di trasmissione è un elemento da minimizzare in un sistema di codifica. Queste considerazioni sono alla base dell'esistenza di codificatori con lunghezze di trama diverse. Il valore più ampiamente utilizzato rimane forse 20 ms.

Come già detto tuttavia, la soluzione ottima consisterebbe in un frame variabile di lunghezza dipendente dal particolare fonema pronunciato. L'impiego di un frame di lunghezza variabile trova applicazione negli schemi di codifica a velocità variabile di cui si parlerà in seguito. Viceversa, nel caso in cui la lunghezza del frame sia costante, diventa importante assicurare una graduale transizione nel passare da un set di coefficienti ad un set successivo. Tale graduale cambiamento è intrinsecamente giustificato dal modello di produzione del segnale vocale che è caratterizzato da una variazione graduale continua dell'assetto articolatorio. Il metodo più semplice per ottenere tale transizione smussata, consiste nell'interpolazione, solitamente lineare, dei parametri spettrali relativi a set adiacenti. Non tutti i set di coefficienti spettrali esaminati prima forniscono le stesse prestazioni da questo punto di vista.

Una misura della bontà delle prestazioni in questo caso è data dalla distorsione spettrale misurata tra lo spettro interpolato e lo spettro effettivo, nel punto di interpolazione. La figura 5.24 [Toh79] riporta l'andamento della distorsione spettrale per diversi tipi di parametri. Si nota che, nell'intervallo di

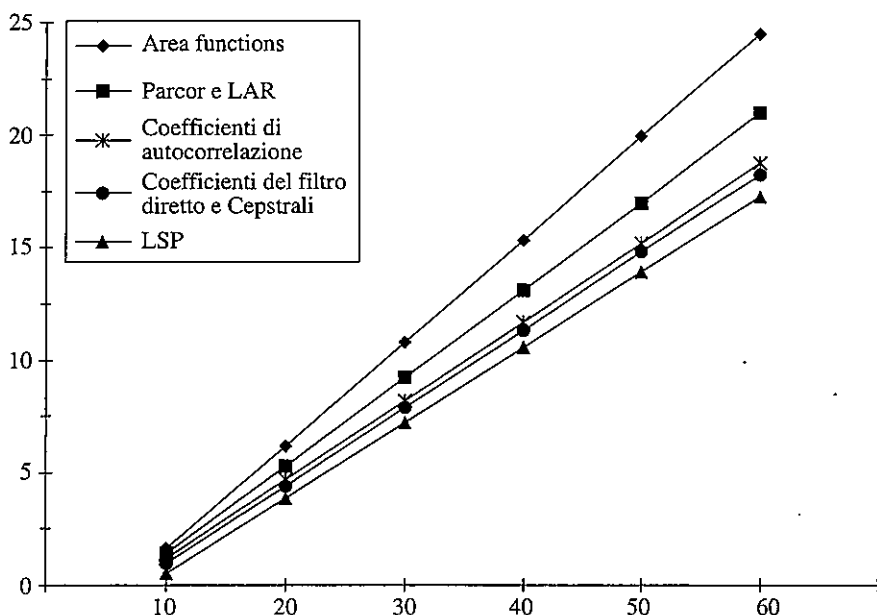



Fig. 5.24 - Distorsione spettrale per diversi set di parametri LPC interpolati.

durate di interesse, la distorsione aumenta al crescere della lunghezza del frame di analisi in quanto le caratteristiche spettrali tendono ad essere maggiormente diverse. Inoltre si nota che i parametri LSP forniscono le prestazioni migliori anche da questo punto di vista in quanto consentono una minore distorsione spettrale.

Questi dati si riferiscono ai valori medi di distorsione calcolati su un certo numero di frasi relative a diversi parlatori e quindi fanno riferimento a prestazioni medie. Tuttavia bisogna osservare che dal punto di vista della qualità del segnale ricostruito, distorsioni locali seppur con probabilità di occorrenza bassa, ma con ampiezza alta, possono comportare risultati di qualità soggettiva insoddisfacenti.

In questo senso l'esigenza di interpolare le caratteristiche spettrali ha validità in generale, ma esistono situazioni particolari in cui può comportare risultati peggiorativi. È il caso di particolari suoni, come gli attacchi, in cui le caratteristiche spettrali cambiano bruscamente e la loro interpolazione comporta distorsioni locali di ampiezza molto superiore ai valori medi. Questo fenomeno è solitamente mitigato negli schemi di codifica parametrica, in quanto una buona quantizzazione del segnale di eccitazione può sopperire ad errori della stima

spettrale. Tuttavia in uno schema specifico, lo standard per codifica Half-rate del GSM [ETSI06.20], proprio per sopperire a tale inconveniente, è stato riservato un bit per segnalare al ricevitore quali frame interpolare e quali no.

5.2.6 Codifica DPCM

Come già più volte ripetuto, nella codifica predittiva ciò che viene trasmesso è la differenza d(n) tra il segnale e la sua stima, che coincide con l'errore di predizione

$$d(n) \equiv e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k) \qquad (5.114)$$

La struttura del codificatore (fig. 5.25) è quindi quella di un filtro numerico che, a partire dai campioni, fornisce l'errore di predizione (prediction-error filter). La relativa funzione di trasferimento (a soli zeri) è pari a

$$C(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \qquad (5.115)$$

Il decodificatore (filtro di ricostruzione) esegue localmente, in funzione dei precedenti campioni del segnale, la stessa stima eseguita in trasmissione e ricostruisce il segnale originale sommando alla stima del campione corrente il valore ricevuto della funzione differenza

$$x(n) = d(n) + \sum_{k=1}^{p} \alpha_k x(n-k) \qquad (5.116)$$

La sua funzione di trasferimento (a soli poli) è, quindi, pari a

$$D(z) = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \qquad (5.117)$$

Si nota che le funzioni di trasferimento del codificatore e del decodificatore sono, ovviamente, l'uno l'inverso dell'altro. In tal modo la funzione di trasferi-

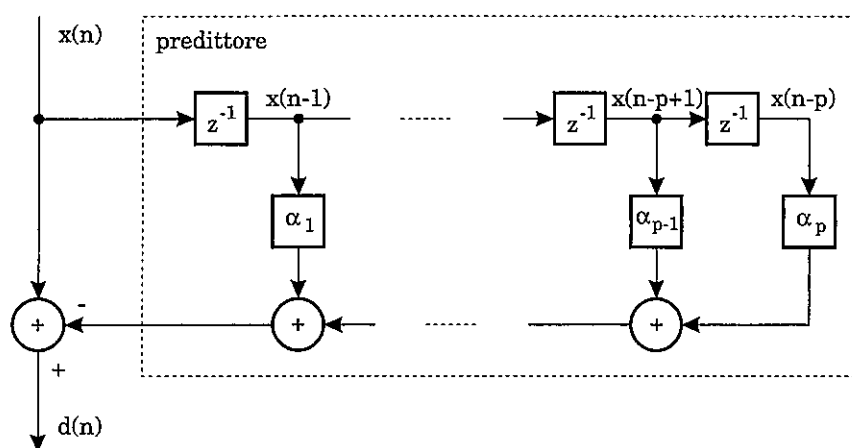



Fig. 5.25 - Struttura del codificatore per quantizzazione differenziale.

mento complessiva, data dal prodotto delle due, è unitaria e, dopo la decodifica, è possibile riottenere il segnale in ingresso al codificatore. Inoltre, ricordando che, nel caso di predizione ottima, il segnale differenza d(n) coincide con l'ingresso v(n) del modello ARX della sorgente e che i coefficienti a si ottengono cambiando di segno a quelli della sorgente stessa, la funzione di trasferimento del decodificatore coincide con quella del modello ARX della sorgente

$$H(z) = \frac{1}{1 + \sum_{k=1}^{q} w_k z^{-k}} \tag{5.118}$$

Ipotizzando in ingresso al filtro ARX del rumore bianco, e quindi con spettro piatto, il modulo dello spettro del segnale prodotto dalla sorgente coincide con il modulo di tale funzione di trasferimento (fig. 5.26). Di conseguenza, avendo il codificatore una funzione di trasferimento che risulta essere l'inverso della funzione densità spettrale di potenza dell'ingresso, la sua uscita, cioè il segnale differenza, ha spettro piatto e quindi risulta essere rumore bianco. Per tale motivo, il codificatore è generalmente indicato anche come "filtro di sbiancamento" o "filtro inverso". La codifica DPCM, quindi, può essere vista come un sistema di identificazione dello spettro del segnale d'ingresso. Infatti, dato lo spettro del segnale P(ω) e quello prodotto dall'ARX

$$\hat{P}(\omega) = \frac{1}{|1 + \sum_{k=1}^{p} \alpha_k e^{-jk\omega}|^2} \tag{5.119}$$

è possibile riottenere le equazioni di Wiener-Hopf minimizzando l'errore spettrale

$$E = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega \tag{5.120}$$

Lo sbiancamento dello spettro del segnale differenza può essere anche dimostrato differentemente. Infatti, la funzione di autocorrelazione del segnale differenza è pari a

$$R_{dd}(i) = E\{ d(n)\, d(n-i) \} \; ; \quad i = 1, 2, \ldots \tag{5.121}$$

Sostituendo in essa l'espressione del predittore

$$d(n-i) = x(n-i) - \hat{x}(n-i) = x(n-i) - \sum_{j=1}^{\infty} \alpha_j\, x(n-i-j) \; ; \quad i = 1, 2, \ldots \tag{5.122}$$

e dato che, per il principio di ortogonalità, risulta

$$E\{ d(n)\, x(n-i) \} = E\{ e(n)\, x(n-i) \} = 0 \; ; \quad i = 1, 2, \ldots \tag{5.123}$$

si ottiene

$$R_{dd}(i) = - \sum_{j=1}^{\infty} \alpha_j\, E[\, d(n)\, x(n-i-j)\,] \; ; \quad i = 1, 2, \ldots \tag{5.124}$$

Riapplicando il principio di ortogonalità, si verifica che la funzione di auto-correlazione risulta essere impulsiva e quindi, nell'ipotesi (essenziale) che il predittore sia di ordine infinito, il segnale differenza è in teoria rumore bianco.

Ovviamente, la predizione del segnale eseguita con predittori di ordine finito (predizione a breve termine) riesce a ricostruire solamente l'andamento medio dello spettro del segnale. Nel caso di segnale vocale, ad esempio, con un

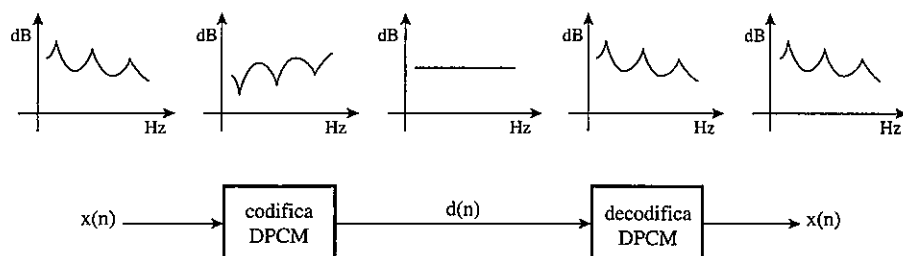



Fig. 5.26 - Relazione tra spettri dei segnali e funzioni di trasferimento in un sistema di codifica differenziale.

predittore di ordine limitato ad esempio, all'ottavo ordine, si riesce a determinare solamente la posizione delle prime quattro formanti del segnale. Rifacendoci al modello discreto della sorgente, in tal modo si ricostruisce solo il contributo del filtro (cioè nell'individuazione della risposta impulsiva). La struttura fine dello spettro, dovuta ad un'eccitazione non bianca, ma periodica, viene persa, ma può essere recuperata tramite differenti algoritmi, come descritto in seguito (predizione a lungo termine). Il non perfetto adattamento del codificatore allo spettro dell'ingresso fa si che lo spettro del segnale differenza sia colorato.

Da quanto detto, nell'ipotesi di predittore con ordine infinito e non considerando gli effetti della quantizzazione del segnale differenza, la codifica predittiva risulta perfettamente reversibile: il segnale ricostruito, cioè, coincide esattamente con l'ingresso. Requisito essenziale per la reversibilità della codifica differenziale è che la stima eseguita dal trasmittente sia analoga a quanto fatto in ricezione (fig. 5.27).

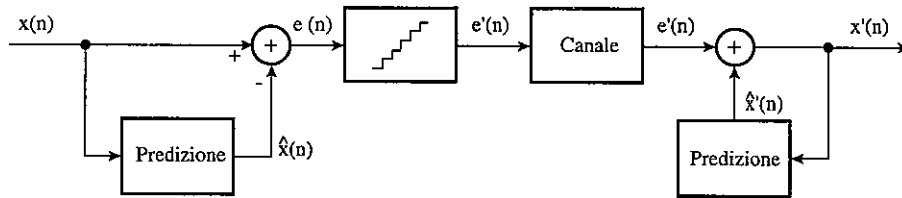
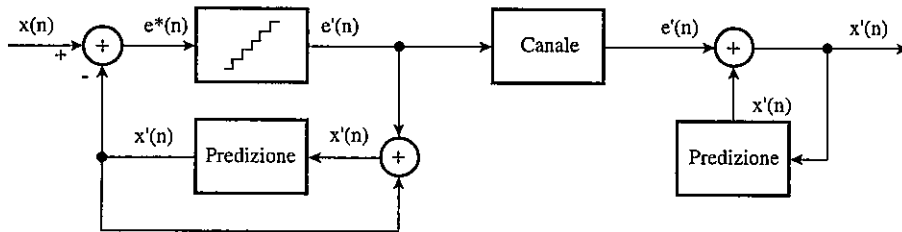
In tale contesto è necessario considerare l'effetto della degradazione dovuta alla quantizzazione del segnale differenza. In decodifica, infatti, il segnale ricostruito dalla stima e dal segnale differenza è inevitabilmente affetto da rumore di quantizzazione. Di conseguenza, anche le successive stime risultano alterate da tale rumore. Qualora in trasmissione la stima venisse fatta direttamente sui campioni del segnale (codifica ad anello aperto), il risultato della predizione risulterebbe differente quanto fatto in ricezione. Il segnale differenza trasmesso, quindi, non permetterebbe di ricostruire fedelmente l'ingresso. Per evitare tale disuniformità, è necessario eseguire la stima anche in trasmissione sulla base di un segnale ricostruito dalle stime precedenti corrette tramite il segnale differenza quantizzato (codifica ad anello chiuso).

Codifica differenziale ad anello aperto

Codifica differenziale ad anello chiuso

Fig. 5.27 - Codifica differenziale ad anello aperto e chiuso.

In ogni caso, il segnale differenza quantizzato porta ad un segnale ricostruito differente dall'originale. Indicando con $x_t(n)$ e $x_r(n)$ l'ingresso e l'uscita del sistema di codifica, con $\hat{x}(n)$ l'uscita del preditore e con d(n) e d'(n) il segnale differenza privo e affetto da rumore di quantizzazione $e_q(n)$, risulta

$$\begin{cases} d(n) = x_t(n) - \hat{x}(n) \\ d'(n) = d(n) + e_q(n) \quad \rightarrow \quad x_r(n) = x_t(n) + e_q(n) \\ x_r(n) = \hat{x}(n) + d'(n) \end{cases} \tag{5.125}$$

Si può, quindi, affermare che l'unico contributo all'errore di codifica per una quantizzazione differenziale è dato dal rumore di quantizzazione.

L'ultimo aspetto che si vuole affrontare riguarda i criteri con i quali fissare la caratteristica di quantizzazione per il segnale differenza. Dato che il segnale differenza risulta essere rumore bianco, la quantizzazione da adottare può essere uniforme. Nell'ipotesi di quantizzazione uniforme e rapporto segnale rumore (e quindi ampiezza dei quanti) pari al caso di quantizzazione non differenziale, la compressione del segnale può avvenire, quindi, solo per una riduzione del numero dei livelli dovuta ad una riduzione degli estremi di

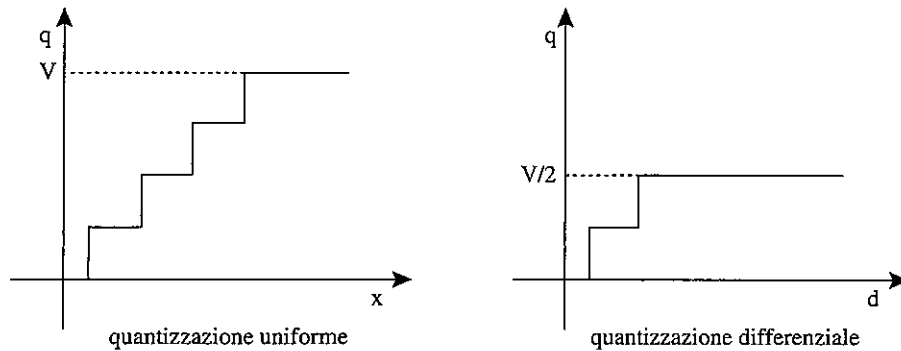



Fig. 5.28 - Riduzione dei livelli del quantizzatore tramite codifica differenziale.

saturazione del quantizzatore (fig. 5.28). Tale riduzione dipende dal guadagno di predizione.

Per determinare il guadagno di predizione è necessario valutare la varianza del segnale differenza

$$\sigma_d^2 = E\{[x(n) - \hat{x}(n)]^2\} \tag{5.126}$$

Nel caso di predizione ottima, tale varianza coincide con il minimo del funzionale d'errore, per cui

$$\sigma_d^2 = \sigma_x^2 - \sum_{k=1}^{p} \alpha_k R(k) \tag{5.127}$$

introducendo l'autocorrelazione normalizzata $\rho(k) = R(k) / \sigma_x^2$, il guadagno di predizione è pari a

$$G = \frac{\sigma_x^2}{\sigma_d^2} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k \rho(k)} \tag{5.128}$$

Il guadagno di predizione, dunque, è funzione dell'efficacia dello stimatore e, quindi, della correlazione del segnale. A tale proposito è necessario considerare che, nel caso del segnale telefonico, essendo questo filtrato passa banda, l'efficacia dello stimatore è ridotta rispetto al caso di segnali filtrati passa basso, per la mancanza delle componenti a frequenza inferiore (più

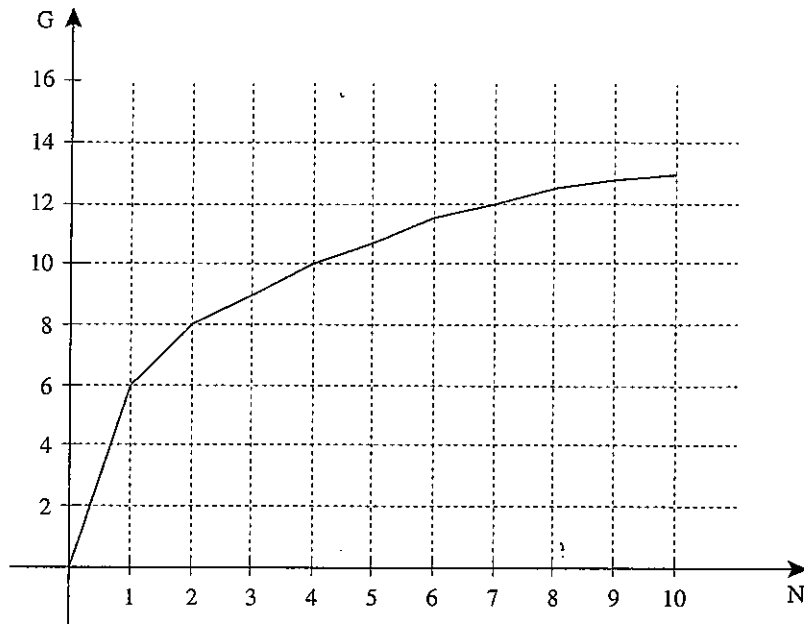



Fig. 5.29 - Guadagno di predizione per predizione adattativa.

fortemente correlate). Inoltre l'efficienza della stima non può essere considerata stazionaria, in quanto risulta maggiore per suoni vocalizzati, caratterizzati da forme d'onda stabili e periodiche, rispetto a suoni non vocalizzati. Fissando la soglia di saturazione sul guadagno di predizione che si ha con suoni vocalizzati, comunque, non si hanno problemi di codifica in quanto, essendo i suoni non vocalizzati gli elementi del segnale ad ampiezza inferiore, questi ultimi non dovrebbero portare a fenomeni di saturazione con la riduzione della dinamica del quantizzatore.

Per segnale vocale, il guadagno di predizione per algoritmi di predizione adattativi in avanti migliora rispetto alla predizione con coefficienti costanti, raggiungendo un massimo di circa 12 dB con predittori di ordine pari all'ottavo (fig. 5.29). Aumentando ulteriormente l'ordine del predittore, non si notano incrementi apprezzabili. Nel caso di predizione all'indietro il guadagno si riduce di circa 1 dB [Jay84]. A parità di rapporto segnale/rumore con la codifica PCM, questo si traduce in una riduzione di due bit per campione, con un segnale differenza esprimibile su 6 bit. In tal modo, il flusso dati si riduce a 48 kb/s, rispetto ai 64 kb/s del logPCM.

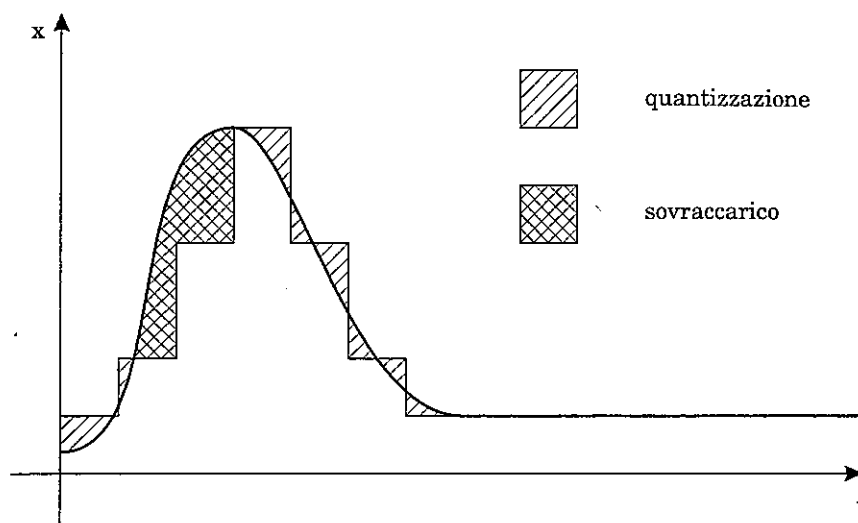



Fig. 5.30 - Errore di sovraccarico.

Per ridurre ulteriormente la componente granulare dell'errore di quantizzazione, è possibile adottare una quantizzazione non uniforme. Infatti, nel caso di predizione non ottima, il segnale differenza non è bianco. Nel caso di segnale vocale, la distribuzione dell'errore di predizione che si osserva è gaussiana, per cui è possibile adottare una quantizzazione non uniforme ottimizzata per tale tipo di distribuzione.

Considerando, infine, gli effetti dell'errore di saturazione nella quantizzazione del segnale differenza, si nota come esso si manifesti come una particolare forma di distorsione del segnale. Infatti, in presenza di rapide variazioni della dinamica dell'ingresso, l'insufficiente correzione apportata dal codificatore alla predizione per effetto della saturazione, comporta l'impossibilità per il segnale decodificato a seguire l'ingresso (rumore di sovraccarico o "slope overload") (fig. 5.30). La presenza di tale tipo di distorsione ha una ripercussione sullo spettro del rumore di quantizzazione, che si arricchisce di componenti a frequenza minore e, quindi, non risulta più né bianco, né completamente scorrelato con l'ingresso.

5.3 CODIFICA ADPCM

Come già accennato, le prestazioni migliori di codificatori adattativi (sia per quanto riguarda la quantizzazione che per quanto riguarda la predizione) si ottengono tramite una loro combinazione (Adaptive Differential Pulse Code Modulation: ADPCM) (fig. 5.31). Il codificatore relativo, applica al segnale differenza ottenuto dalla predizione una quantizzazione in grado di adattarsi alla dinamica dell'errore. Nella raccomandazione ITU-T G.721, il codificatore ADPCM utilizza, sia per il quantizzatore che per il predittore, un adattamento all'indietro, in modo tale che non vi è nessuna informazione aggiuntiva scambiata tra trasmittente e ricevente, al di fuori del segnale differenza.

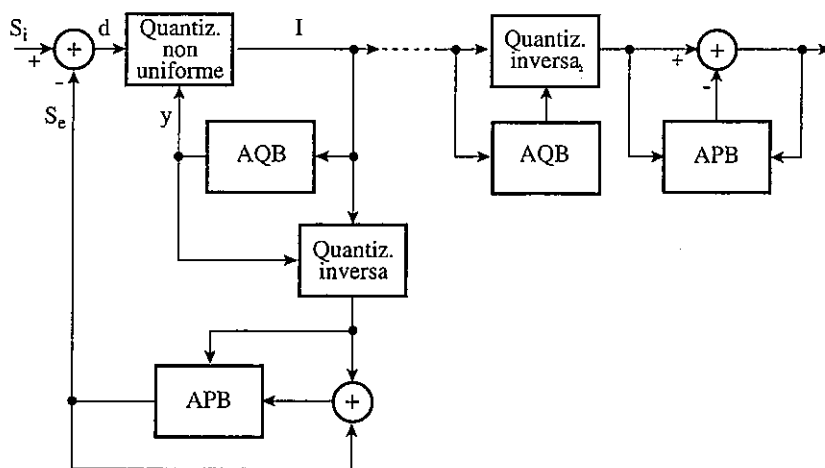



Fig. 5.31 - Codifica ADPCM.

Dato che il codificatore risulta essere, in realtà, un transcoder, trasformando codici LogPCM in codici ADPCM e viceversa, il primo passo è quello di ottenere dalla codifica logaritmica una rappresentazione lineare $s_l(k)$. Ottenuta questa, viene poi ricavata il segnale differenza d(k) dalla stima $s_e(k)$ dell'ingresso, come

$$d(k) = s_l(k) - s_e(k) \tag{5.129}$$

Per quanto riguarda la quantizzazione AQB del segnale differenza (fig. 5.32), la caratteristica del quantizzatore non è uniforme, ma risulta essere

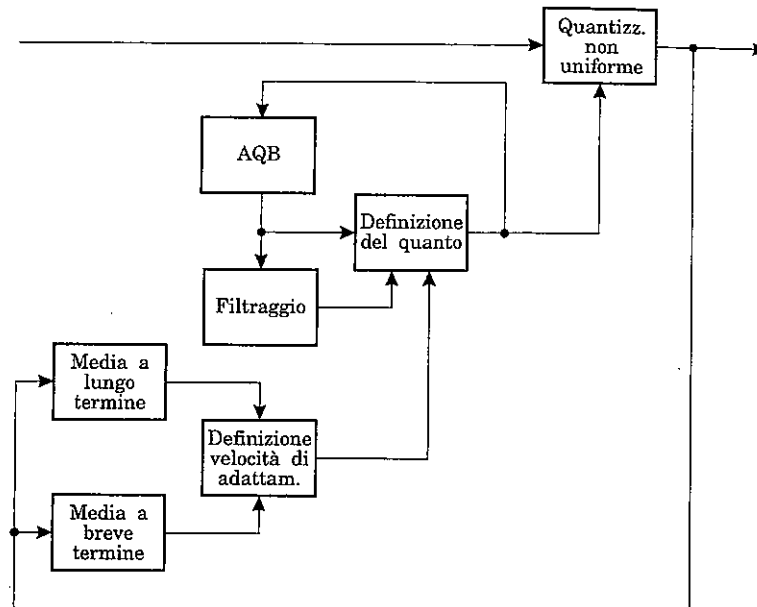



Fig. 5.32 - Quantizzazione adattativa per codifica ADPCM.

simile alla caratteristica ottima per segnali gaussiani [Jay76]. Dato che si utilizza come ingresso del quantizzatore il logaritmo del segnale differenza, l'adattamento del passo di quantizzazione

$$q(n) = \frac{x(n)}{\Delta(n)} \tag{5.130}$$

si trasforma dalla divisione per il Δ alla sottrazione di un fattore di scala di quantizzazione y(k). Il risultato I(k)

$$| I(k) | \leftarrow \log_2 | d(k) | - y(k) \tag{5.131}$$

rappresenta l'uscita del codificatore ADPCM ed ha segno pari al segno dei campioni e modulo dato dalla tabella 5.3 (fig. 5.34).

Codifica $\log_2 \mid d(k) \mid - y(k)$	$\mid I(k) \mid$	Decodifica $\log_2 \mid d(k) \mid - y(k)$
[3.16, + ∞)	7	3.34
[2.78, 3.16)	6	2.95
[2.42, 2.78)	5	2.59
[2.04, 2.42)	4	2.23
[1.58, 2.04)	3	1.81
[0.96, 1.58)	2	1.29
[- 0.005, 0.96)	1	0.53
(- ∞, - 0.005)	0	- 1.05

Tab. 5.3 - Codifica dell'uscita.

Il passo di quantizzazione y(k) è dato dalla combinazione di due componenti $y_u(k)$ e $y_l(k)$, ottimizzate per segnali che variano rapidamente (es.: voce) o lentamente (es.: trasmissioni dati in banda vocale con modulazione PSK). La componente $y_u(k)$ è ottenuta tramite una legge di tipo "robust"

$$\Delta(n) \;=\; M(\mid I(n-1) \mid)\;\; \Delta^{1-\beta}(n-1) \qquad (5.132)$$

Dato che il passo di quantizzazione y(k) è espresso logaritmicamente, il moltiplicatore M si trasforma nella somma di un termine W, funzione dell'uscita I(k), secondo la relazione

$$y_u(k) = 2^{-5}\, W[I(k)] + [1 - 2^{-5}]\, y(k)$$
$$1.06 \leq y_u(k) \leq 10.00 \qquad (5.133)$$

La legge di corrispondenza tra W ed I è data dalla seguente tabella

$\mid I(k) \mid$	7	6	5	4	3	2	1	0
W[I(k)]	69.25	21.25	11.50	6.12	3.12	1.69	0.25	-0.75

Tab. 5.4 - Moltiplicatori logaritmici.

a cui corrispondono i seguenti moltiplicatori $M = 2^{\left(2^{-5} W\right)}$

	8	7	6	5	4	3	2	1
M_i	4.482	1.585	1.283	1.142	1.070	1.037	1.006	0.984

Tab. 5.5 - Moltiplicatori.

La componente $y_l(k)$ è ottenuta filtrando la componente $y_u(k)$ come

$$y_l(k) = [1 - 2^{-6}]\, y_l(k-1) + 2^{-6}\, y_u(k) \qquad (5.134)$$

Il contributo delle due componenti del passo di quantizzazione avviene secondo un fattore di velocità $a_l(k)$ tramite la seguente relazione

$$y(k) = a_l(k)\, y_u(k-1) + [\, 1 - a_l(k)\,]\, y_l(k-1) \qquad (5.135)$$

$a_l(k)$ assume valori compresi tra zero ed uno. Per quanto riguarda la legge secondo la quale è aggiornato $a_l(k)$, innanzitutto esso è sottoposto ad un decadimento esponenziale del tipo

$$a(k) = [1 - 2^{-4}]\, a(k-1) \qquad (5.136)$$

Inoltre, il suo valore dipende dalla dinamica del segnale, ricavata da due medie a breve e lungo termine $d_{ms}(k)$ e $d_{ml}(k)$ ottenute come

$$d_{ms}(k) = [1 - 2^{-5}]\, d_{ms}(k-1) + 2^{-5}\, F[I(k)]$$

$$d_{ms}(k) = [1 - 2^{-7}]\, d_{ms}(k-1) + 2^{-7}\, F[I(k)] \qquad (5.137)$$

dove la funzione F[I(k)] è descritta nella seguente tabella

$\mid I(k) \mid$	7	6	5	4	3	2	1	0
F[I(k)]	7	3	1	1	1	0	0	0

Tab. 5.6 - Correzione della dinamica.

Nel caso in cui il canale diventi poco attivo ($y(k) < 3$) o la differenza tra medie a breve ($d_{ms}(k)$) e lungo termine ($d_{ml}(k)$) del segnale d'uscita diventi eccessiva, cioè

$$| d_{ms}(k) - d_{ml}(k) | \geq 2^{-3} d_{ml}(k) \tag{5.138}$$

la legge di variazione del coefficiente di velocità viene modificata in

$$\begin{cases} a_l(k) = \left[1 - 2^{-4} \right] a(k-1) + 2^3 \\ a_l(k) \leq 1 \end{cases} \tag{5.139}$$

tendendo ad 1 e facendo così prendere il sopravvento alla componente $y_u(k)$ del fattore di scala. Altrimenti $a_l(k)$ decade verso lo zero, facendo così prendere il sopravvento alla componente $y_l(k)$.

Passando alla predizione APB, essa opera sulla versione quantizzata della funzione differenza ricavata dall'uscita I(k)

$$d_q(k) = \log_2^{-1} \left\{ [\log_2 | d(k) | - y(k)] + y(k) \right\} \tag{5.140}$$

Il predittore non è di tipo ARX, ma ad un filtro IIR a due poli, corrispondente ad un predittore del secondo ordine, è associato un filtro FIR a 6 zeri (fig. 5.33). Ciò è motivato principalmente dal rischio di instabilità del predittore a solo poli nel caso di errori di trasmissione o consistenti errori di quantizzazione, ma anche per un migliore adattamento allo spettro di segnali vocali e dati [Jay76]. La stima è data dalla somma dei contributi del filtro FIR ed IIR

$$s_e(k) = s_{ez}(k) + s_{ep}(k) \tag{5.141}$$

I coefficienti di entrambi i filtri vengono adattati secondo l'algoritmo del gradiente sfruttando la cross-correlazione di polarità tra il segnale d'ingresso e di errore ed un decadimento esponenziale (per mitigare la presenza di errori)

$$\alpha_k^{(n+1)} = \lambda\, \alpha_k^{(n)} + \mu \operatorname{sgn}[e(n)] \operatorname{sgn}[x(n-k)] \tag{5.142}$$

con $\lambda = 255/256$ e $\mu = 1/128$. Per quanto riguarda il filtro IIR, la sua uscita costituisce la stima $s_e(k)$ del segnale. Essa viene calcolata in funzione dell'uscita del filtro FIR e del segnale ricostruito agli intervalli precedenti

$$s_r(k) = s_e(k - i) + d_q(k - i) \tag{5.143}$$

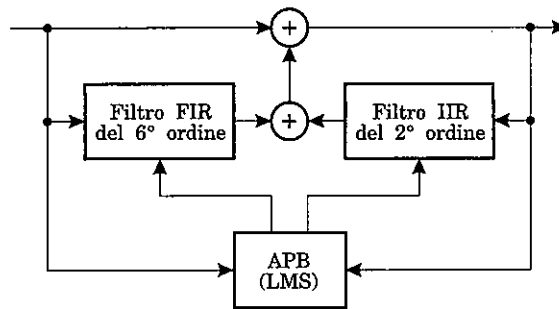



Fig. 5.33 - Predizione adattativa per codifica ADPCM.

secondo la relazione

$$s_e(k) = \sum_{i=1}^{2} a_i(k-1)\, s_r(k-i) + s_{ez}(k) \tag{5.144}$$

L'algoritmo del gradiente per l'aggiornamento dei coefficienti del filtro IIR è applicato ad una funzione p(k), ottenuta eliminando dal segnale ricostruito il contributo del filtro stesso

$$p(k) = s_{ez}(k) + d_q(k) \tag{5.145}$$

L'aggiornamento dei coefficienti del filtro IIR è poi ottenuto come

$$\begin{cases} a_2(k) = \left[1-2^{7}\right] a(k-1) + 2^{-7}\, \mathrm{sgn}[p(k)]\, \mathrm{sgn}[p(k-1)] - f[a_1(k-1)]\, \mathrm{sgn}[p(k)]\, \mathrm{sgn}[p(k-1)] \\ |\, a_2(k)\, | \ \le \ 0.75 \end{cases}$$

$$\begin{cases} a_1(k) = [\, 1 - 2^{-8}\,]\, a_1(k-1) + 3 \cdot 2^{-8}\, \mathrm{sgn}[\, p(k)\,]\, \mathrm{sgn}[\, p(k-1)\,] \\ |\, a_1(k)\, | \ \le \ 1 - 2^{-4} - a_2(k) \end{cases} \tag{5.146}$$

dove

$$\begin{cases} f(a_i) = \begin{cases} 4\,a_1 & \text{se } |a_1| \leq ½ \\ 2\,\text{sgn}(a_1) & \text{se } |a_1| > ½ \end{cases} \\ \text{sgn}(0) = +1 \end{cases} \tag{5.147}$$

Per quanto riguarda l'uscita $s_{ez}(k)$ del filtro FIR, essa è data da

$$s_{ez}(k) = \sum_{i=1}^{6} b_i(k-1)\, d_q(k-i) \tag{5.148}$$

Per l'aggiornamento dei coefficienti $b_i(k)$, non sarebbe possibile applicare l'LMS, che è stato sviluppato per un sistema ARX. Se si ripetono, però, gli stessi passi che hanno portato all'LMS nel caso di sistema MA, si ottiene una relazione del tipo

$$\alpha_k^{(n+1)} = \lambda\, \alpha_k^{(n)} + \mu\, \text{sgn}[\, x(n)\,]\, \text{sgn}[\, x(n-k)\,] \tag{5.149}$$

dove si nota che l'aggiornamento è funzione del gradiente tra il segnale x(n) e il segnale al nodo i-esimo del filtro x(n - i). Nel nostro caso, considerando il segnale differenza $d_q(k)$, si ottiene

$$\begin{cases} b_i(k) = [1 - 2^{-8}]\, b_i(k-1) + 2^{-7}\, \text{sgn}[\, d_q(k)\,]\, \text{sgn}[\, d_q(k-i)\,]\,; \quad i = 1, 2, \ldots, 6 \\ -2 \leq |\, b_i(k)\,| \leq +2 \end{cases} \tag{5.150}$$

Con tale codifica, a parità di rapporto segnale/rumore, è guadagno di predizione è tale da ridurre il numero di bit per campione a 4 bit, con un throughput di 32 kb/s contro i 48 kb/s del DPCM.

5.4 SAGOMATURA DELLO SPETTRO DELL'ERRORE

Come già visto, nella codifica differenziale l'errore di codifica coincide con il rumore di quantizzazione. La distribuzione uniforme dello spettro del rumore di quantizzazione, però, non risulta ottimale dal punto di vista della percezione. In tal modo, infatti, il rapporto segnale rumore non risulterebbe costante in frequenza. Nel caso di segnale vocale, ad esempio, questo

migliorerebbe in corrispondenza delle componenti armoniche a più alta energia, cioè le formanti, e viceversa. Sebbene la parte dello spettro relativo alle formante è effettivamente la più importante dal punto di vista della percezione, tale andamento del rapporto SN non sfrutta gli effetti di mascheramento esistenti nell'apparato uditivo. Questi permetterebbero un aumento del rumore di quantizzazione in corrispondenza delle componenti ad energia maggiore dello spettro, senza una degradazione percettibile del segnale.

Riprendendo lo schema di un codificatore differenziale ad anello chiuso (fig. 5.35), questo può essere interpretato come quello di un codificatore PCM che lavora su una versione "compressa" in frequenza dell'ingresso, ottenuta tramite un filtro con una funzione di trasferimento complementare allo spettro del segnale. La "decompressione" è, poi, eseguita in decodifica. Per analizzare l'effetto di tali trasformazioni sullo spettro del rumore di quantizzazione, è opportuno riscrivere l'espressione del codificatore differenziale ad anello chiuso esplicitando il termine relativo al rumore

$$d(n) = x(n) - \hat{x}(n) = x(n) - \sum_{j=1}^{p} \alpha_j x'(n-j)$$

$$= x(n) - \sum_{j=1}^{p} \alpha_j x(n-j) - \sum_{j=1}^{p} \alpha_j e_q(n-j) \tag{5.151}$$

Si nota che in una configurazione ad anello chiuso, il rumore è assoggettato alle stesse trasformazioni di compressione e decompressione che vengono eseguite sul segnale e, quindi, il suo spettro in uscita torna ad essere piatto. Un modo per non avere in uscita un rumore con spettro piatto (noise shaping) può essere quello di conservare l'effetto di sagomatura del codificatore, eliminando, però, il suo filtraggio inverso in codifica. Per il codificatore, dunque, dovrebbe valere una legge del tipo

$$d(n) = x(n) - \sum_{j=1}^{p} \alpha_j x(n-j) + e_q(n) \tag{5.152}$$

che vuol dire passare ad uno schema di codifica ad anello aperto. In tal caso, sommando al segnale compresso un rumore di quantizzazione a spettro piatto,

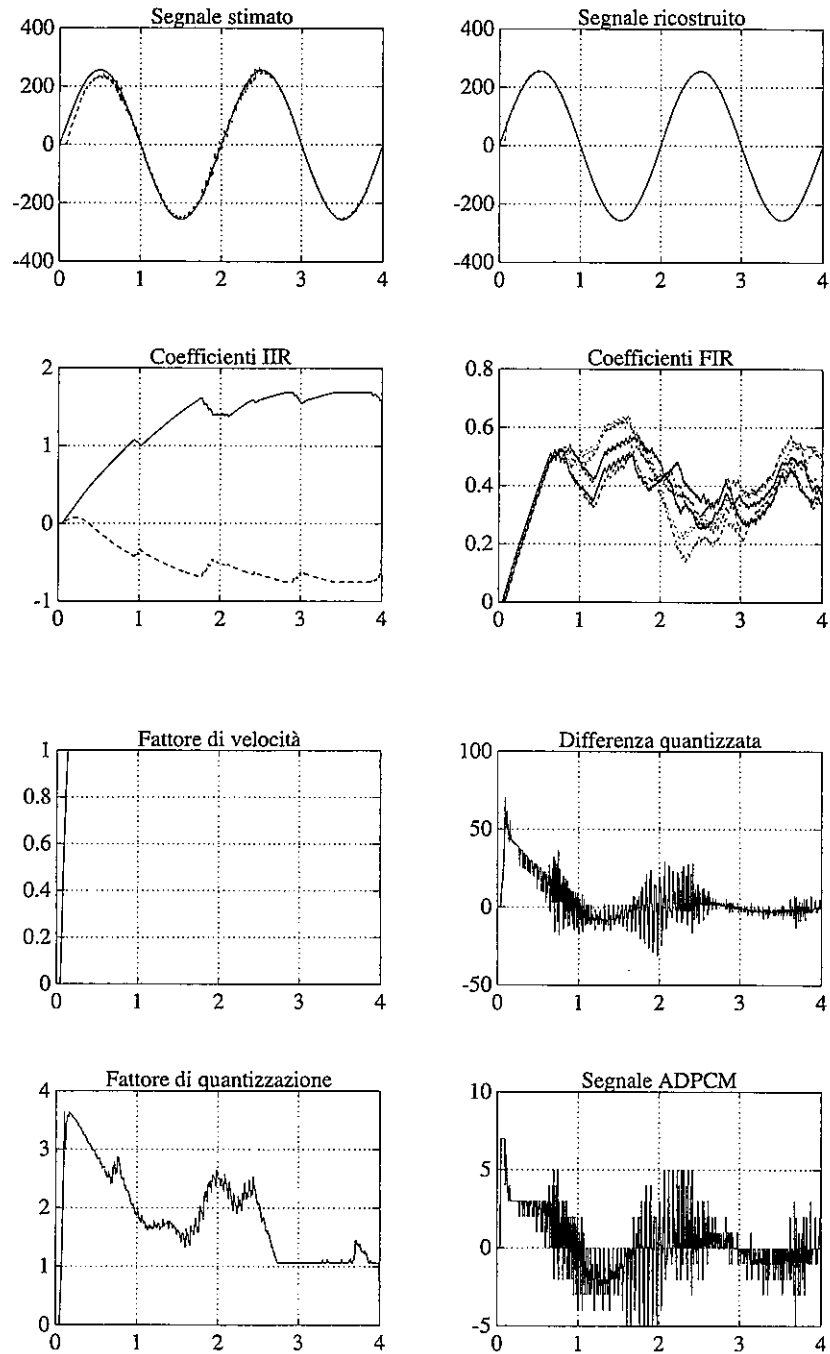



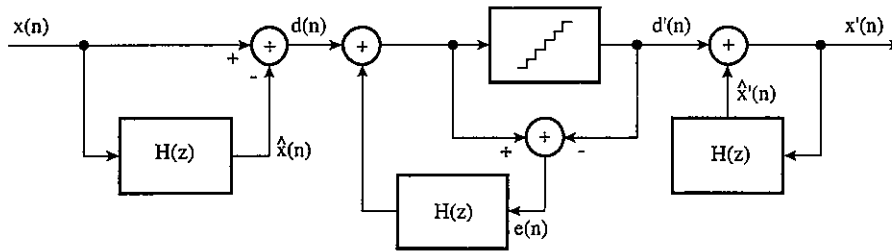
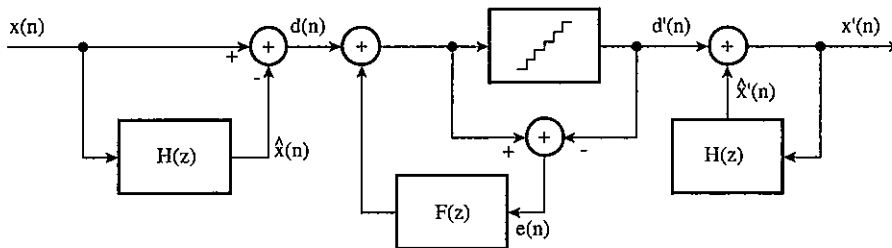
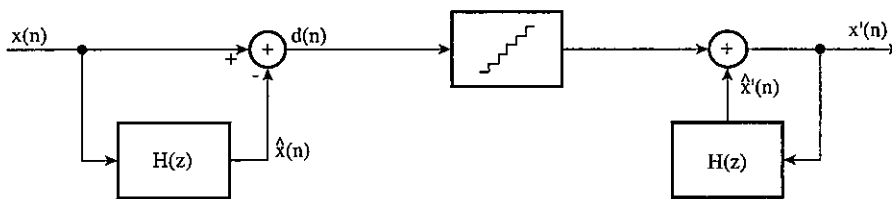
Fig. 5.34 - Segnale derivante da una codifica ADPCM.

Codifica differenziale ad anello chiuso

Noise Feedback Coding

Codifica differenziale ad anello aperto

Fig. 5.35 - Noise Feedback Coding come caso particolare di codifiche differenziali.

questo verrà poi sagomato in frequenza in fase di decodifica, comprimendolo nelle zone ad energia minore ed esaltandolo nelle zone ad energia maggiore (fig. 5.36). In questo modo si ottiene la voluta normalizzazione del rapporto segnale rumore.

Ovviamente, tra tali due estremi (filtraggio del rumore con l'inverso dello spettro del segnale o non filtraggio) esistono soluzioni intermedie nelle quali il rumore di quantizzazione viene filtrato indipendentemente dallo spettro del segnale. Tali codifiche vengono indicate come Noise Feed-back Coding (NFC).

L'indipendenza dallo spettro del segnale del filtro per la sagomatura in frequenza del rumore (noise-weighting filter) introduce un ulteriore parametro

per l'ottimizzazione della qualità di codifica. È da notare come l'incremento di complessità del codificatore avvenga esclusivamente in trasmissione, in quanto il decodificatore non è interessato dalla procedura di sagomatura.

Calcolando la potenza del segnale di errore in caso di sagomatura si ottiene un leggero incremento rispetto al caso di spettro uniforme. Anche se la potenza di rumore globale risulta essere aumentata, gli effetti di mascheramento da parte dell'apparato uditivo comportano comunque un miglioramento della qualità percepita.

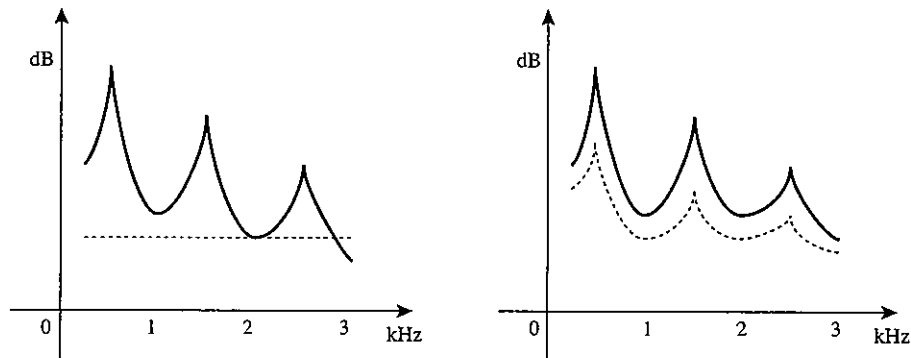



Fig. 5.36 - Sagomatura del rumore di quantizzazione.

5.5 MODULAZIONE DELTA

5.5.1 Sovracampionamento

Nella trattazione relativa alla conversione A/D e D/A con quantizzazione uniforme è stato considerato un campionamento a frequenza f_s pari a quella di Nyquist. Come risultato di tale campionamento si ha, innanzitutto, la periodicizzazione dello spettro del segnale a multipli della frequenza di campionamento. D'altro canto si genera un rumore di quantizzazione, caratterizzato da uno spettro approssimativamente piatto. La potenza del rumore di quantizzazione che cade nella banda del segnale è pari a $\Delta^2 / 12$, con densità spettrale di potenza costante e pari a

$$\frac{\Delta^2}{12\, f_s} \tag{5.153}$$

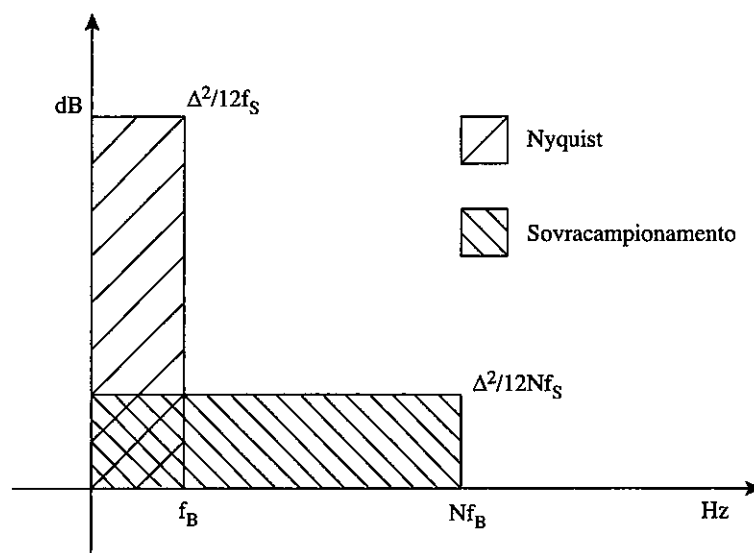



Fig. 5.37 - Densità spettrale di potenza per campionamento alla frequenza di Nyquist e sovracampionamento.

Sia lo spettro del segnale che la densità spettrale del rumore dipendono, dunque, dalla frequenza di campionamento. Se questa viene aumentata, ad esempio campionando ad una frequenza ($N f_s$) multipla di quella di Nyquist (sovracampionamento), un primo risultato che si ottiene è che la rappresentazione in frequenza del segnale campionato presenterà repliche del segnale nell'intorno di una frequenza di campionamento maggiore e quindi più spaziate tra loro. Di conseguenza, è possibile adottare filtri con zone di transizione più ampie, riducendo le distorsioni in banda introdotti dagli stessi.

Il vantaggio maggiore del sovracampionamento, però, si ha dal punto di vista della riduzione del rumore di quantizzazione. Essendo il livello della densità spettrale di potenza inversamente proporzionale alla frequenza di campionamento, essa si riduce proporzionalmente al fattore di sovracampionamento N adottato (fig. 5.37). Filtrando il segnale al di fuori della banda del segnale, la potenza di rumore che cade nella banda stessa viene, quindi, ridotta. Il rumore di quantizzazione che si ottiene sovracampionando è quello tipico di convertitori ad un maggiore numero di bit di quanto effettivamente utilizzato nel convertitore. Viceversa, a parità di rumore di quantizzazione, il sovracampionamento permette di ridurre il numero di bit del convertitore. Dato

che la semplificazione della parte analogica del convertitore compensa l'incremento di complessità dovuto alla maggiore frequenza di lavoro, la conversione tramite sovracampionamento risulta tecnologicamente meno critica e quindi più economici.

Trascurando per il momento un'eventuale riduzione di bit per campione, al fine di mantenere costante il flusso numerico prodotto è necessario eseguire due conversioni della frequenza di campionamento [Cro83](fig. 5.38). Innanzitutto è necessario ridurre il numero di campioni al secondo dopo la conversione A/D (decimazione) e poi risovracampionare prima della conversione D/A (interpolazione). Concettualmente, la decimazione consiste in un nuovo campionamento del segnale alla frequenza di Nyquist. Essendo il ricampionamento in realtà eseguito eliminando un certo numero di campioni, come

$$y(m) = w(N \times m) \tag{5.154}$$

è necessario che il fattore di sovracampionamento N sia un intero. Il segnale da decimare w(n) è ottenuto dall'ingresso sovracampionato sottoposto ad un filtraggio numerico passa basso, al fine di evitare fenomeni di aliasing

$$w(n) = \sum_{k=-\infty}^{\infty} h(k)\; x(n-k) \tag{5.155}$$

per cui

$$y(m) = \sum_{k=-\infty}^{\infty} h(k)\; x(N \times m - k) = \sum_{n=-\infty}^{\infty} h(N \times m - n)\; x(n) \tag{5.156}$$

Per quanto riguarda la conversione D/A, invece, è necessario aumentare la frequenza di campionamento prima della trasformazione in analogico. Questo può essere ottenuto inserendo tra campioni consecutivi del segnale un numero di campioni nulli pari al fattore di sovracampionamento

$$w(m) = \begin{cases} x\left(\frac{m}{N}\right), & m = \pm N, \pm 2N, \ldots \\ 0, & \text{altrimenti} \end{cases} \tag{5.157}$$

Tali campioni fittizi vengono poi corretti effettuando un'interpolazione con i campioni reali, tramite un filtraggio numerico passa basso con frequenza di taglio pari a quella di sovracampionamento

$$y(m) = \sum_{k=-\infty}^{\infty} h(m-k)\, x\left(\frac{k}{N}\right) = \sum_{n=-\infty}^{\infty} h(m - N \times n)\, x(n) \qquad (5.158)$$

In tal modo vengono eliminate le repliche introdotte dal sovracampionamento stesso. Per quanto riguarda l'implementazione dei filtri FIR coinvolti nella conversione di frequenza di campionamento, si rimanda in [Appendice D].

5.5.2 Modulazione delta lineare ed adattativa

Oltre ai benefici tecnologici precedentemente evidenziati, il sovracampionamento ha risvolti positivi anche per quanto riguarda la codifica del segnale. Sovracampionando il segnale fino ad arrivare a frequenze di campionamento di ordini di grandezza superiori a quella di Nyquist (es.: 10 MHz per segnale audio) si ottengono due tipi di vantaggi. Da un lato, a parità di rumore si ha una riduzione del numero di bit necessari per la quantizzazione, Dall'altro si aumenta la correlazione tra campioni. Infatti la funzione di autocorrelazione del segnale campionato è data dal campionamento della funzione di autocorrelazione del segnale continuo. Aumentando la frequenza di campionamento, se ne infittiscono i campioni e, di conseguenza, si riduce l'ampiezza delle variazioni tra suoi campioni adiacenti. Un aumento della correlazione del segnale comporta un aumento del guadagno di predizione. Infatti, riprendendo l'espressione del guadagno

$$G = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k\, \rho(k)} \qquad (5.159)$$

e riscrivendola per p=1, con il coefficiente di predizione pari a $\alpha_1 = R(1) / R(0) = \rho(1)$ si ottiene

$$G = \frac{1}{1 - \rho^2(1)} \qquad (5.160)$$

da cui si osserva che, per $\rho \rightarrow 1$, il guadagno tende all'infinito, con la conseguente riduzione di dinamica del segnale differenza. La limitazione ad un

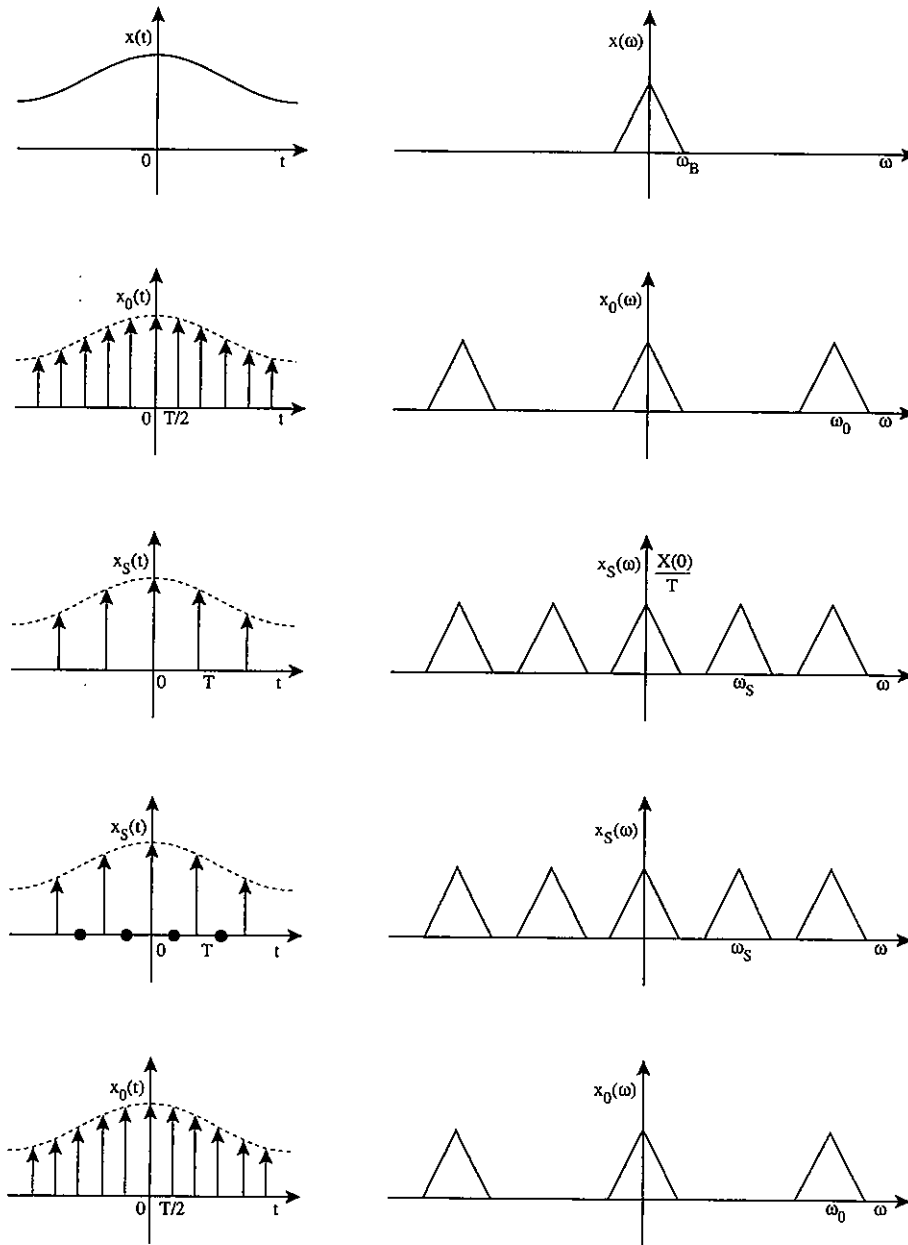



Fig. 5.38 - Effetti di un sovracampionamento per N=2 nel dominio del tempo e della frequenza.

predittore del primo ordine è giustificata dall'elevata efficienza del predittore stesso ed in linea con la semplicità del codificatore (fig. 5.39).

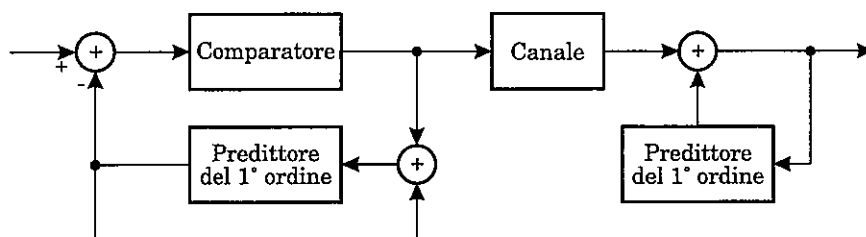



Fig. 5.39 - Struttura di un codificatore DM lineare.

L'azione combinata di tali due fattori porta ad una riduzione progressiva del numero di livelli di quantizzazione all'aumentare della frequenza di campionamento, fino ad arrivare a rendere possibile la codifica del segnale su di un solo bit. In tale codifica, detta modulazione delta (Delta Modulation: DM), non è necessario utilizzare un convertitore A/D per la codifica del segnale differenza, in quanto è necessario rilevare solamente il suo segno. Ciò riduce la componente analogica del codificatore ad un semplice comparatore. La codifica e decodifica DM con un predittore del primo ordine avviene, dunque, con i seguenti passi

$$\begin{cases} \hat{x}(n) = \alpha\, x(n-1) \\ d(n) = \mathrm{sgn}[\, x(n) - \hat{x}(n)\,] \times \Delta \\ x(n) = \hat{x}(n) + d(n) \end{cases} \tag{5.161}$$

Siccome $\alpha \approx 1$, la codifica DM può essere semplificata in

$$\begin{cases} d(n) = \mathrm{sgn}[\, x(n) - x(n-1)\,] \times \Delta \\ x(n) = x(n-1) + d(n) \end{cases} \tag{5.162}$$

da cui si vede che la codifica si basa sulle differenze tra campioni adiacenti e la decodifica si ottiene semplicemente integrando la funzione d'errore.

Per le caratteristiche dell'errore di codifica nella DM, valgono tutte le considerazioni fatte a proposito del DPCM. Per quanto riguarda l'ampiezza del

quanto, affinché la codifica non dia origine a slope-overload, esso deve soddisfare la seguente relazione

$$\frac{\Delta}{T} \geq \max \left| \frac{d\,x(t)}{dt} \right| \tag{5.163}$$

Con tale modulazione, però, nel caso di segnale d'ingresso costante si introduce una forma ineliminabile di idle channel noise (fig. 5.40). Infatti, in corrispondenza di un simile ingresso, l'uscita dal codificatore non può rimanere costantemente nulla, ma oscillerà con un'ampiezza pari ad un quanto.

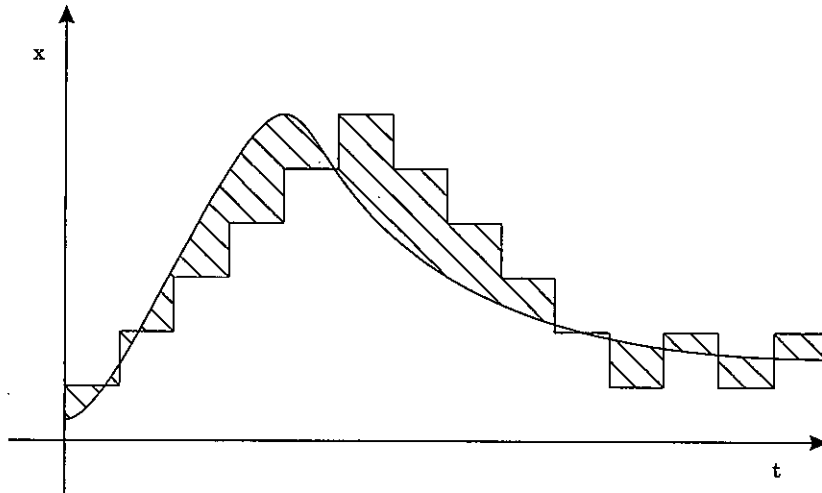



Fig. 5.40 - Rumore di sovraccarico e granulare per modulazione delta lineare.

Entrambi i problemi di slope-overload e di idle channel noise possono essere attenuati ricorrendo, anche per la modulazione delta come per il DPMC, ad una quantizzazione adattativa (Adaptive DM: ADM).

Nelle applicazioni correnti, la velocità dei dati trasmessi con una modulazione delta, se non abbinata ad altre tecniche di compressione, è analoga a quella del logPCM (64 kb/s). Utilizzando una quantizzazione adattativa si può ridurre il throughput del codificatore a 32 kb/s.

5.5.3 Convertitori sigma-delta

Dalla modulazione delta deriva la modulazione sigma-delta, che permette di migliorare ulteriormente il rapporto segnale-rumore della conversione tramite tecniche di trasferimento in frequenza della potenza del rumore di campionamento (noise shaping). La chiave di tale tecnica risiede nella riorganizzazione dei blocchi funzionali che compongono un modulatore delta (fig. 5.41). In particolare, il primo passo è il trasferimento dell'integratore in ricezione (posto prima del filtro d'uscita), in ingresso al codificatore, data la linearità dello stesso. Il secondo passo consiste nel trasferimento di entrambi gli integratori (quello di ingresso e quello per il calcolo della funzione d'errore) all'interno dell'anello del quantizzatore. Pur non variando la funzione globale di tale codificatore, in tal modo si rende la funzione di trasferimento del rumore dipendente dalla frequenza. Infatti, non considerando l'effetto del rumore di quantizzazione, per il segnale si ha la seguente funzione di trasferimento

$$Y(s) = \frac{X(s) - Y(s)}{s}$$

$$\frac{Y(s)}{X(s)} = \frac{\frac{1}{s}}{1 + \frac{1}{s}} = \frac{1}{s+1} \tag{5.164}$$

mentre per il rumore, con ingresso nullo, si ha

$$Y(s) = \frac{-Y(s)}{s} + N(s)$$

$$\frac{Y(s)}{N(s)} = \frac{1}{1 + \frac{1}{s}} = \frac{s}{s+1} \tag{5.165}$$

Dall'analisi ti queste funzioni di trasferimento si nota come il segnale risulti filtrato passa basso, mentre il rumore passa alto (fig. 5.42). Dato che le componenti a frequenza maggiore del rumore saranno eliminate dal filtro d'interpolazione d'uscita, in tal modo si riduce ulteriormente la potenza del rumore di quantizzazione che cade nella banda del segnale.

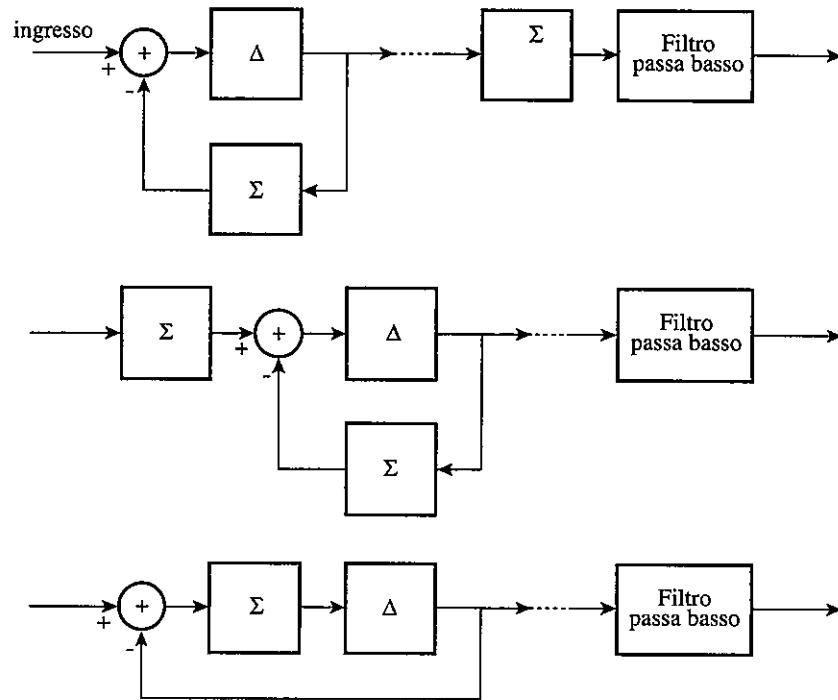



Fig. 5.41 - Modulazione sigma delta.

5.5.4 Conversione DM-PCM

La semplicità realizzativa di un codificatore DM è essenziale per il raggiungimento delle elevate frequenze di campionamento richieste. D'altra parte la riduzione del rumore di quantizzazione permesso da convertitori sigma-delta e l'eliminazione delle distorsioni introdotte da convertitori A/D e D/A di codificatori PCM lineari, fanno si che anche per la codifica PCM lineare si preferisca eseguire una doppia conversione analogico-DM e DM-PCM.

Il problema della conversione DM-PCM si riconduce al problema di trasformazione della frequenza di campionamento di un segnale numerico, già introdotto a proposito del sovracampionamento. La conversione DM-PCM, in particolare, consiste nella decimazione del segnale DM, preceduto da un filtraggio numerico passa basso. Durante tale filtraggio viene eseguito anche l'incremento di bit per campione con i quali viene espresso il codice.

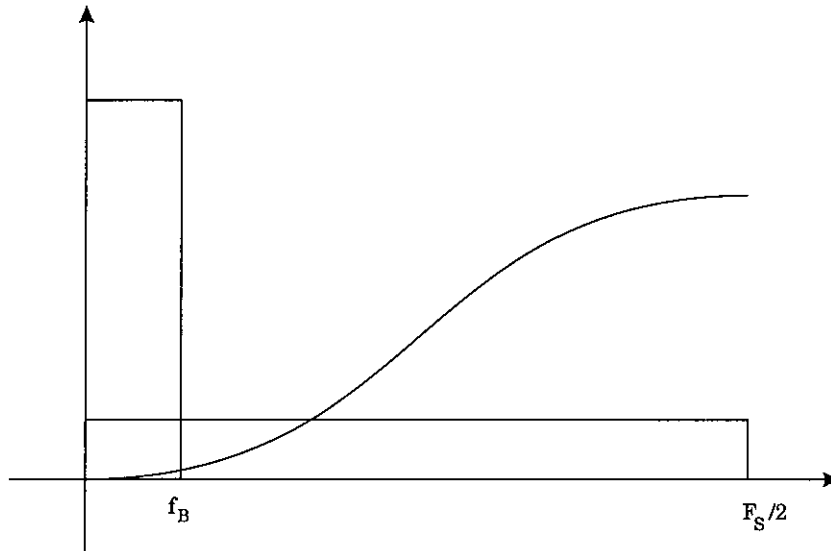



Fig. 5.42 - Densità spettrale di potenza per codificatori PCM, DM e Sigma-Delta.

Dato, però, che i campioni DM possono assumere solamente due valori, la conversione può essere ottenuta semplicemente tramite un contatore con un numero di bit pari a quelli della codifica PCM, il clock del quale è pari a quello del flusso DM e controllo up/down ottenuto dal valore dei bit trasmessi.

5.6 PREDIZIONE A LUNGO TERMINE E APC

Il tipo di predizione precedentemente esposta viene definita predizione a breve termine ($p \leq 12$) ed è valida per qualsiasi tipo di segnale. Il problema della predizione può essere affrontato in modo sostanzialmente differente in presenza di segnali periodici, come per i suoni vocalizzati nel caso della voce. È possibile, infatti, adottare una predizione a lungo termine che tenti di determinare il valore del campione prossimo in funzione del corrispondente campione del periodo precedente (fig. 5.43). Ciò si attua tramite una predizione del tipo

$$\hat{x}(n) = \beta \, x(n-N) \tag{5.166}$$

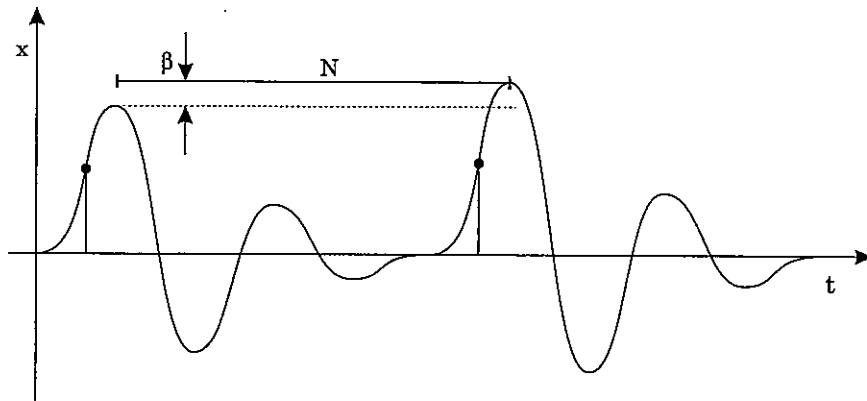



Fig. 5.43 - Predizione a lungo termine.

dove N ($20 \leq N \leq 120$) coincide con il numero di campioni presenti in un periodo, mentre β rappresenta le variazioni di ampiezza tra periodi adiacenti. Per l'individuazione del valore della costante N (pitch prediction) è possibile, ad esempio, ricorrere all'analisi della funzione di autocorrelazione del segnale. Considerando segnali periodici stazionari, infatti, la funzione di autocorrelazione, stimata da un blocco di M campioni come

$$R(k) = \sum_{n=0}^{M-1} x(n)\, x(n+k) \tag{5.167}$$

è una funzione periodica con lo stesso periodo del segnale. Per segnali periodici quasi stazionari, invece, essa assume periodicamente valori molto prossimi a R(0). Analizzando l'andamento della R(k), quindi, è possibile ottenere il periodo del segnale N come distanza del primo massimo relativo dall'origine, calcolata in numero di campioni. Tale analisi è semplificata se, oltre ad eseguire sul blocco di campioni un filtraggio passa basso al di sotto del kHz, essi vengono anche "tagliati" in ampiezza, lasciando passare solo la parte del segnale di livello maggiore.

Per il calcolo del fattore di guadagno β, si ricorre alla minimizzazione rispetto a questo parametro dell'errore quadratico

$$E = \sum_{n} [\, x(n) - \hat{x}(n-N) \,]^2 \tag{5.168}$$

Sostituendo l'espressione dello stimatore a lungo termine ed imponendo l'annullamento della $\partial E / \partial \beta$, si ricava che il valore ottimo del fattore di guadagno è pari a

$$\beta = \frac{E[\,x(n)\,x(n-N)\,]}{E[\,x^2(n-N)\,]} \tag{5.169}$$

La predizione a lungo termine è solitamente associata una predizione a breve termine e la combinazione delle due è detta Adaptive Predictive Coding (APC) (fig. 5.44). In tal caso la stima del segnale d'ingresso è ottenuta come

$$\hat{x}(n) = \sum_{k=1}^{p} \alpha_k\, x(n-k) + \beta \left[x(n-N) - \sum_{k=1}^{p} \alpha_k\, x(n-k-N) \right] . \tag{5.170}$$

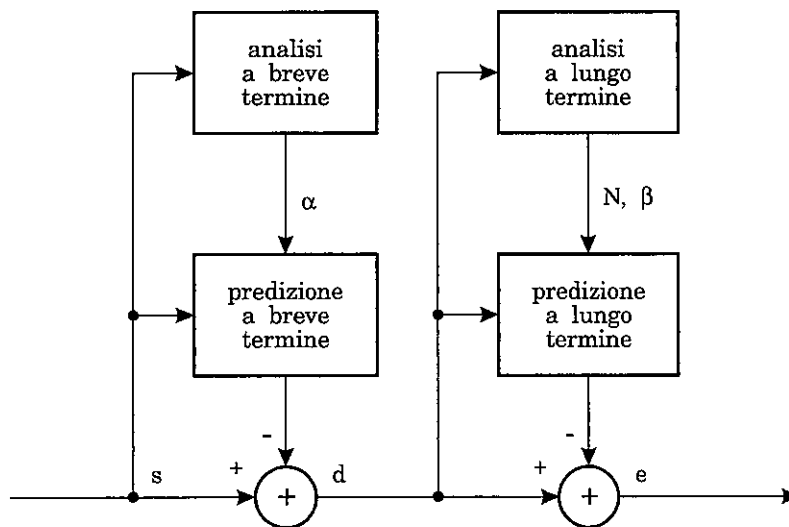



Fig. 5.44 - Codifica APC.

La giustificazione della combinazione di predizione a breve e lungo termine deriva dal fatto che il residuo della prima nel caso di segnali vocalizzati è composto da un treno periodico di impulsi, essendo legato all'eccitazione del modello ARX della sorgente (fig. 5.45). Sull'ampiezza di tali impulsi è fissata

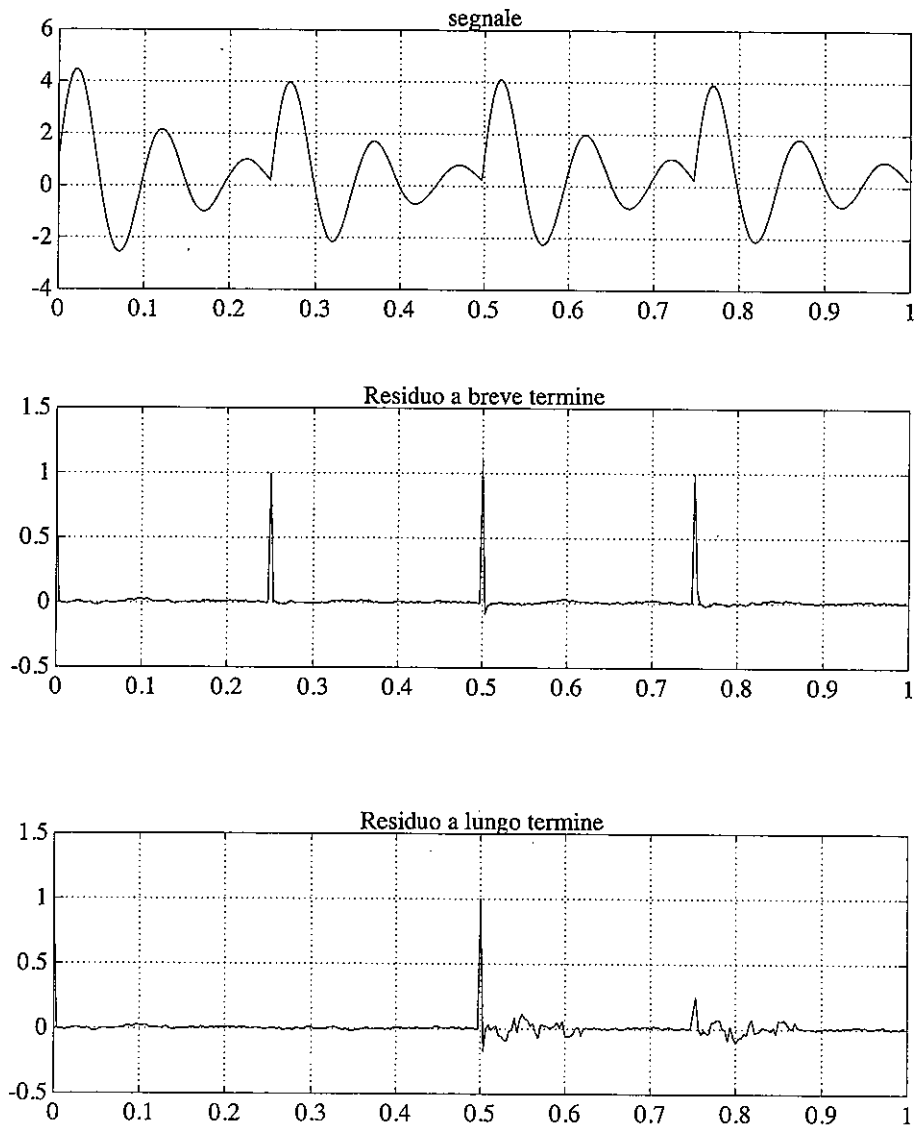



Fig. 5.45 - Residui di predizione a breve e a lungo termine.

l'ampiezza del livello di saturazione del quantizzatore e, quindi, il numero di bit necessari per la codifica. Tali limiti, quindi, risultano sovradimensionati rispetto alla media dei campioni, normalmente di ampiezza modesta. Eliminando tali impulsi, quindi, è possibile comprimere ulteriormente il flusso numerico.

Guardando il codificatore APC come uno stimatore dello spettro del segnale d'ingresso e, quindi, dell'individuazione del modello digitale della sorgente, si nota come la componente a breve termine del predittore sia in grado di ricostruire l'inviluppo dello spettro e quindi il contributo del filtro ARX, mentre la componente a lungo termine è in grado di ricostruire la variazione più fine dello stesso, legata all'eccitazione.

Dal punto di vista dell'implementazione, dato che la determinazione congiunta delle tre variabili N, β e α_k risulta essere difficoltosa, tipicamente viene ottimizzata dapprima la predizione a breve e poi quella a lungo termine. Inoltre, il predittore a lungo termine non può essere che a blocchi. Volendo, infatti, trasmettere per ogni campione i due parametri N e β necessari per la stima, tale codifica comporterebbe un'espansione del flusso generato dalla sorgente. Dal punto di vista dell'efficacia della predizione, però, questo non rappresenta un limite in quanto, limitando la dimensione dei blocchi all'interno di un periodo del segnale (es.: 40 campioni), i due parametri dovrebbero mantenersi sufficientemente costanti all'interno del blocco stesso. D'altra parte, essendo l'APC utilizzata per il raggiungimento di livelli di compressione elevati, anche il predittore a breve termine è tipicamente a blocchi.

6

CODIFICA PER MODELLI

6.1 CODIFICA PER MODELLI NEL DOMINIO DELLA FREQUENZA E NEL TEMPO

Le codifiche di forma d'onda precedentemente discusse si prefiggono di trasmettere al ricevente l'informazione sull'andamento nel tempo del segnale da codificare, al fine di permetterne la ricostruzione. Nella codifica per modelli, detta anche codifica per analisi e sintesi o parametrica, quello che si vuole fare è di mettere in grado il ricevente di generare localmente il segnale in base ad informazioni sulle caratteristiche della sorgente. Essendo la codifica legata al modello prescelto della sorgente stessa, tali codificatori perdono la caratteristica general purpose dei codificatori di forma d'onda fin qui discussi. Nel caso di segnale vocale sono noti come vocoder (Voice Coder).

Il vantaggio di tale tecnica risulta evidente pensando, ad esempio, alla codifica dell'uscita di un generatore di forme d'onda sinusoidali. In tal caso, invece di trasmettere l'infinita serie dei campioni del segnale, sarebbe sufficiente trasmettere solamente i parametri che identificano il generatore (ampiezza, frequenza e fase iniziale). Con tali informazioni il ricevente sarebbe in grado di ricostruire una replica perfettamente identica all'ingresso del codificatore.

Nel caso della codifica del segnale audio, ovviamente, il problema è di più difficile soluzione, sia per la non stazionarietà che per la complessità della sorgente. Il problema della variazione nel tempo delle caratteristiche del segnale si risolve limitando l'identificazione dei parametri della sorgente ad intervalli temporali entro i quali la stessa può essere approssimata come stazionaria. Trasmessa l'informazione relativa ad un intervallo, l'analisi si ripeterà periodicamente per gli intervalli successivi.

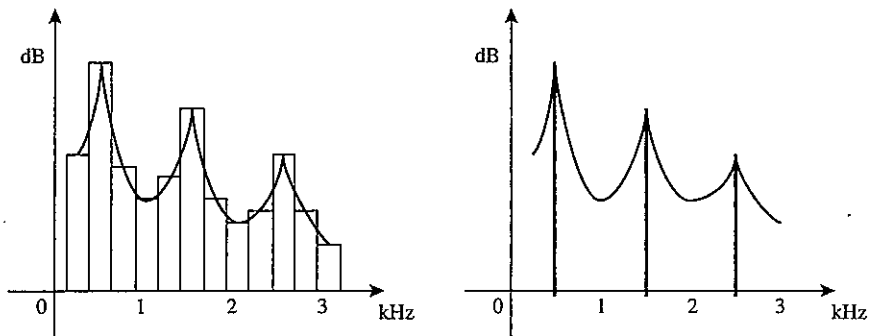



Fig. 6.1 - Vocoder di canale e per formanti.

Per l'identificazione della sorgente, invece, si possono seguire strade differenti. Una possibilità è quella di determinare il modulo dello spettro del segnale (ignorando l'informazione relativa alla fase), per poi procedere alla ricostruzione del segnale tramite generatori sinusoidali. Per un fissato intervallo di tempo, l'ampiezza del segnale prodotto da ciascun generatore è determinabile tramite opportuni filtri passabanda (vocoder di canale) (fig. 6.1).

Utilizzando, ad esempio, 16 sottobande con intervalli di 20 ms e codificando la potenza su 8 bit, il flusso numerico generato risulterebbe di 6400 bit/s. Prestazioni migliori dal punto della riduzione del flusso numerico possono essere ottenute codificando la posizione e l'ampiezza delle sole formanti (vocoder per formanti) (fig. 6.1). I vocoder che operano nel dominio della frequenza, pur se caratterizzati da elevati rapporti di compressione, soffrono di una cattiva qualità del segnale generato.

È possibile realizzare una differente codifica per modelli operando l'identificazione della sorgente nel dominio del tempo (fig. 6.2). In tal caso, la sorgente viene identificata in funzione dei coefficienti del suo modello AR e dell'ingresso. Per la rilevazione di tali grandezze possono essere utilizzati gli algoritmi sviluppati a proposito delle codifiche predittive, in quanto è stato mostrato come:

- dalla struttura del predittore è possibile risalire alla struttura del modello AR della sorgente;
- il residuo di predizione fornisce l'eccitazione del filtro.

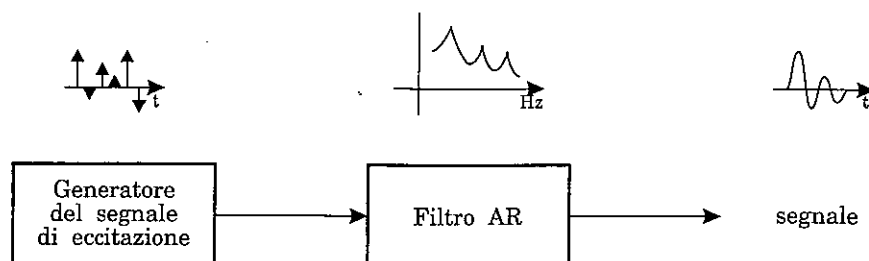



Fig. 6.2 - Codifica per modelli nel dominio del tempo.

L'inconveniente di tale impostazione deriva dal fatto che, anche trascurando il flusso richiesto per la trasmissione dei coefficienti del filtro, per la codifica di forma d'onda dell'eccitazione verrebbe generato un flusso non minore di quello dell'ADPCM.

Una possibile soluzione a questo problema è quella di sottocampionare la sequenza di campioni dell'eccitazione, in considerazione delle caratteristiche spettrali piatte di tale segnale. È questo il caso della tecnica di codifica RELP (Residual Excited Linear Prediction) [Un75] in cui il segnale di eccitazione residuo è limitato in banda, solitamente a 1 kHz e cioè un quarto della banda effettiva, quindi sottocampionato e quantizzato scalarmente o vettorialmente. Al ricevitore il segnale di eccitazione a banda piena è ricostruito con tecniche di ribaltamento spettrale utilizzando il solo contributo della banda da 0 a 1kHz. Sebbene questa tecnica consenta forti rapporti di compressione, la qualità fornita non è molto buona ed il suo impiego è stato molto limitato.

Una seconda soluzione è quella di adottare un modello anche per la codifica dell'eccitazione. Questa strada ha portato allo sviluppo della tecnica di codifica del vocoder LPC ampiamente utilizzata in passato in quanto, nelle versioni più sofisticate, consente velocità di trasmissione anche di poche centinaia di bit/s. Il limite di tale tecnica rimane nel modello di eccitazione molto semplificato che pone un limite intrinseco alla qualità ottenibile.

Il grosso passo in avanti dal punto di vista della qualità si è ottenuto con le tecniche di codifica di analisi per sintesi (ABS). Tali tecniche superano i problemi del vocoder LPC, in quanto non viene utilizzato un modello specifico per il segnale di eccitazione, e consentono ottima qualità con forti rapporti di compressione, in considerazione di due elementi essenziali:

- il segnale di eccitazione è calcolato al trasmettitore utilizzando come criterio la minimizzazione dell'errore del segnale vocale ricostruito, da cui il nome di analisi per sintesi. Questo elemento determina l'esigenza di un decodificatore locale anche nel trasmettitore;
- la minimizzazione dell'errore è realizzata considerando un modello percettivo, seppur semplificato. Tale modello consente di tenere in conto delle trasformazioni operate dall'apparato uditivo.

I codificatori di analisi per sintesi (ABS) si distinguono in tre principali classi e si differenziano essenzialmente per la forma del segnale di eccitazione utilizzato:

- RPE (Regular Pulse Excited). In questo caso il segnale di eccitazione è costituito da un treno di impulsi sottocampionati, di ampiezza e fase opportune, ricavato direttamente dal segnale residuo.
- MPLPC (Multipulse LPC). Il segnale di eccitazione è costituito da un certo numero di impulsi di posizione ed ampiezza opportune.
- CELP (Codebook Excited Linear Prediction). Il segnale di eccitazione è selezionato da una collezione di possibili segnali memorizzati in una tabella.

6.2 LINEAR PREDICTIVE CODING

6.2.1 Generalità

Nella codifica per modelli, il segnale viene generato in ricezione tramite informazioni sul modello della sorgente e sulla sua eccitazione come

$$X(z) = H(z)\, V(z) \tag{6.1}$$

Come già detto, per quanto riguarda la modellizzazione della sorgente questo avviene tramite un sistema lineare auto regressivo con funzione di trasferimento a soli poli del tipo

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{6.2}$$

Questa funzione di trasferimento è soddisfacente solamente per suoni vocalizzati, dato che per le nasali e le fricative sarebbe necessario una funzione di trasferimento a poli e zeri. L'aumento dell'ordine p del sistema, comunque, permette di raggiungere l'approssimazione voluta. Gli α_k, che sono le incognite dell'identificazione, possono essere ricavati tramite algoritmi predittivi (da cui il nome di Linear Predictive Coding [Mar76]). Dato che con questa codifica si è interessati ai rapporti di compressione più spinti, le tecniche utilizzate sono generalmente a blocchi [Appendice C].

L'LPC si distingue per la tecnica di codifica utilizzata per l'eccitazione v(n) del filtro AR (fig. 6.3). Se si utilizzasse il residuo di predizione, il flusso generato sarebbe non inferiore a quello delle codifiche di forma d'onda. Nel caso di segnale vocale, però, anche per l'eccitazione può essere utilizzata una codifica per modelli. Infatti, classificato il segnale come vocalizzato o non vocalizzato, essa può essere ottenuta nel primo caso a partire da un generatore periodico di impulsi, mentre nel secondo caso da un generatore di rumore.

La codifica dell'eccitazione, quindi, richiede innanzitutto l'identifica zione del tipo di generatore da adottare (periodico o meno). Nel caso di suoni vocalizzati, poi, è necessario identificarne anche il periodo. Inoltre, normalmente viene anche esplicitato il guadagno G del filtro, in modo da lavorare su di un'eccitazione normalizzata in ampiezza. Il segnale è, infine, ricostruito come

$$x(n) = \sum_{k=1}^{p} \alpha_k \, x(n-k) + G \, v(n) \tag{6.3}$$

La selezione tra suoni vocalizzati e non vocalizzati e, per quest'ultimi, l'individuazione della fondamentale è uno degli aspetti più complessi della codifica LPC. Per quanto riguarda l'eccitazione dei suoni vocalizzati, l'individuazione corretta della fondamentale (pitch) del treno di impulsi che funge da eccitazione per il modello è di particolare importanza in quanto, pur se non presente nel segnale d'uscita (filtrato passa banda), tramite le sue armoniche determina le variazioni a grana fine dello spettro, con un notevole impatto sulla qualità percepita.

Una tecnica utilizzabile per risolvere entrambi i problemi di selezione tra suoni vocalizzati e non vocalizzati ed individuazione del pitch è l'utilizzo della funzione di auto-correlazione. Come già più volte accennato, per segnali perfet-

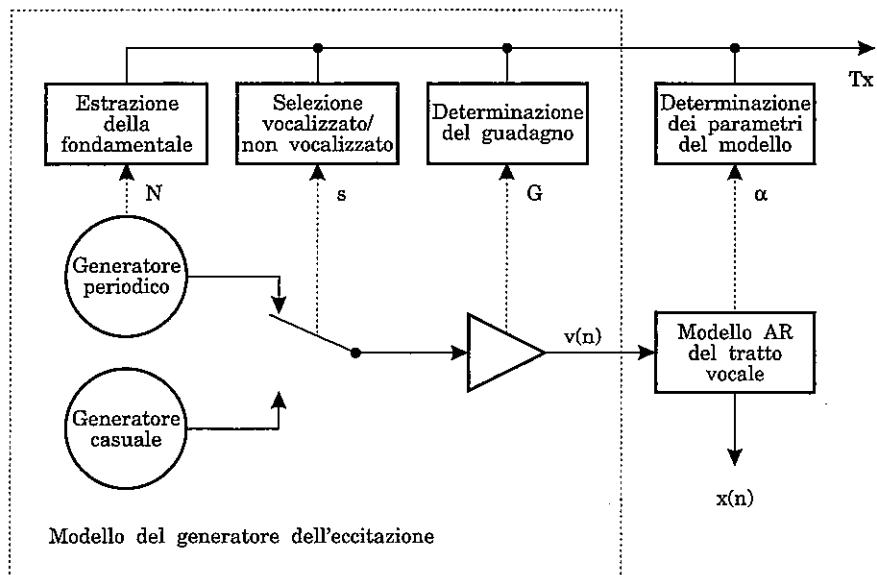



Fig. 6.3 - Codifica LPC.

tamente periodici la funzione d'autocorrelazione riassume periodicamente il suo valore nell'origine. Il segnale può, quindi, essere ritenuto periodico se l'autocorrelazione presenta periodicamente picchi superiori ad una certa soglia (es.: ampiezza pari ad almeno 0.5 R(0)); in tal caso la distanza tra due massimi della funzione di auto-correlazione fornisce anche il periodo N del segnale.

Una stima semplificata della fondamentale può essere ottenuta rinunciando al calcolo dell'autocorrelazione tramite l'Average Magnitude Difference Function (AMDF), definita come

$$D(k) = \sum_{n=1}^{N-1} | x(n) - x(n+k) | ; \quad k = 0, 1, 2, \ldots \tag{6.4}$$

Dato che a distanza di un periodo i campioni dovrebbero mostrare ampiezze simili, l'AMDF permette di individuare la fondamentale del segnale tramite la posizione dei suoi minimi.

Il vantaggio dell'AMDF risiede nella riduzione della complessità computazionale rispetto al calcolo dell'autocorrelazione, grazie alla riduzione del numero di moltiplicazioni richieste. C'è da dire, però, che gli attuali

processori per l'elaborazione dei segnali permettono il calcolo di prodotti con accumulo con le stesse prestazioni con le quali vengono eseguiti operazioni meno complesse, per cui i vantaggi dell'AMDF rispetto all'uso della funzione di autocorrelazione vengono a cadere. Altra grandezza utilizzata per distinguere tra fonemi vocalizzati e non è la misura degli attraversamenti per lo zero (Zero Crossing), definita come

$$\frac{1}{N}\sum_{n=1}^{N}\frac{|\,\mathrm{sgn}[s(n)]-\mathrm{sgn}[s(n-1)]\,|}{2} \tag{6.5}$$

Per quanto riguarda il guadagno G del sistema, nell'ipotesi di identificazione perfetta dei coefficienti del modello AR, G sarebbe calcolabile come rapporto tra il segnale scelto come eccitazione e la funzione d'errore

$$G=\frac{e(n)}{v(n)} \tag{6.6}$$

A causa degli errori di predizione, tale procedimento non è affidabile ed è preferibile confrontare l'energia dei due segnali come

$$G^2=\frac{\sum_{i=1}^{n} e(i)^2}{\sum_{i=1}^{n} v(n)^2} \tag{6.7}$$

Riepilogando, l'eccitazione del filtro lineare viene modellizzata tramite un selettore tra fonemi vocalizzati e non, più due parametri che identificano il periodo e l'ampiezza dell'eccitazione. Dal punto di vista della compressione, quindi, la codifica LPC permette di ottenere flussi estremamente ridotti (throughput fino ai 2.4 kb/s). La qualità ottenuta, però, non è elevata, data la forte semplificazione delle caratteristiche dell'eccitazione.

6.2.2 DoD LPC-10

Un esempio di CoDec LPC è fornito dallo standard governativo statunitense LPC-10 (figg. 6.4 - 6.5). L'ingresso del codificatore sono segmenti

di segnale della durata di 22.5 ms. Dato un campionamento ad 8 kHz, ciò si traduce in blocchi di 180 campioni.

Per la generazione dei parametri LPC, viene utilizzato il metodo della covarianza su sottoblocchi di 130 campioni [Appendice C]. Tali sottoblocchi sono allineati all'inizio del periodo del segnale, in modo da poter eseguire un'analisi sincrona con il periodo stesso, con una riduzione dell'errore di codifica. L'ordine del predittore utilizzato è pari al 10° i suoni vocalizzati e del 4° per i non vocalizzati.

La rilevazione del periodo della fondamentale avviene tramite AMDF sul segnale filtrato passa-basso nell'intervallo 51.3-400 Hz. Per la selezione tra suoni vocalizzati e non vocalizzati si sfruttano le informazioni relative a:

- energia nella banda inferiore dello spettro;
- rapporto tra il massimo e minimo dell'AMDF;
- conteggio degli attraversamenti per lo zero.

Le informazioni sul periodo della fondamentale e sulla selezione tra vocalizzati e non vocalizzati vengono poi corretti tramite tecniche di programmazione lineare.

L'informazione sul guadagno viene, infine, ricavata tramite il calcolo dell'RMS. La codifica avviene secondo la seguente tabella

Parametro	Vocalizzati	Non vocalizzati	Commenti
selezione	1	1	
periodo	6	6	
energia	5	5	
k_1	5	5	LAR
k_2	5	5	LAR
k_3	5	5	lineare
k_4	5	5	lineare
k_5	4		lineare
k_6	4		lineare
k_7	4		lineare
k_8	4		lineare
k_9	3		lineare
k_{10}	2		lineare

Tab. 6.1 - Componenti del flusso LPC10.

Sommando un bit di sincronizzazione (alternativamente 1 o 0) e 21 bit per la protezione dagli errori (nelle trame non vocalizzate), si ottiene un totale di 54 bit per blocco, con un flusso di 2400 bit/s.

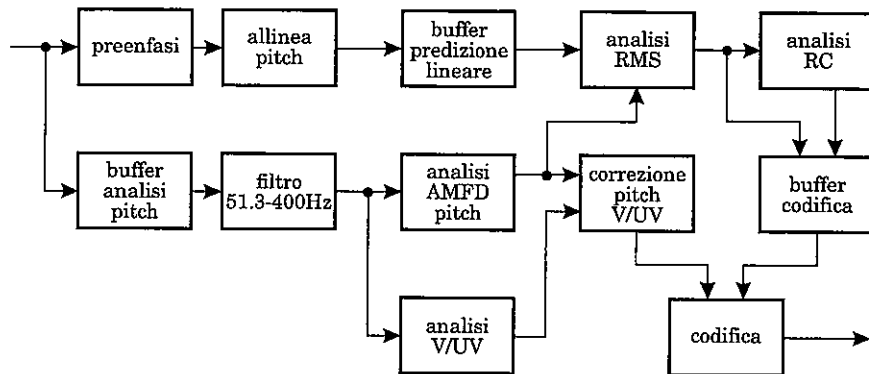



Fig. 6.4 - Struttura del codificatore LPC-10.

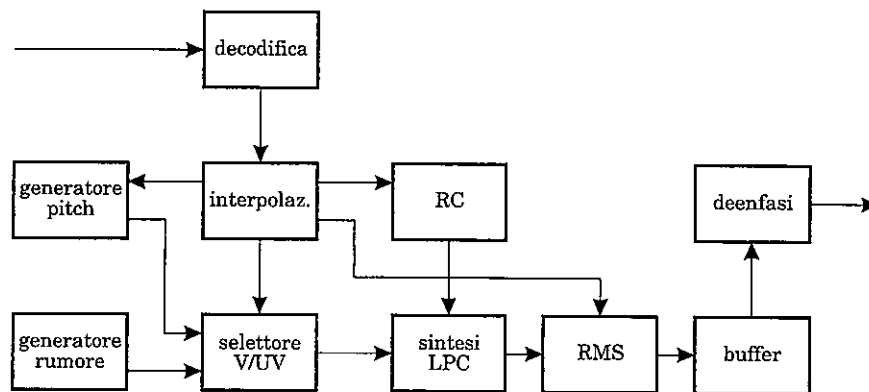



Fig. 6.5 - Struttura del decodificatore LPC-10.

6.3 CODIFICA RPE

6.3.1 Regular-Pulse Excitation

Evitando di utilizzare un modello per il segnale di eccitazione, per un vocoder nel dominio del tempo questo può essere ricavato direttamente dal

residuo di predizione. Tale affermazione può essere verificata ricordando la legge di un vocoder LPC

$$x(n) = G\,v(n) + \sum_{k=1}^{p} \alpha_k\, x(n-k) \tag{6.8}$$

e quella del residuo di predizione a breve termine

$$x(n) = e(n) + \sum_{k=1}^{p} \alpha_k\, x(n-k) \tag{6.9}$$

Dal punto di vista della struttura del codificatore, utilizzando tale tipo di eccitazione, si passerebbe dal sistema ad anello aperto dell'LPC, ad uno retroazionato, noto come codificatore multi-pulse (fig. 6.6). Le tecniche che derivano da tale impostazione hanno, da un lato, il vantaggio di una qualità migliore e, dall'altro, risultano essere applicabili a sorgenti qualsiasi (cioè non esclusivamente alla voce).

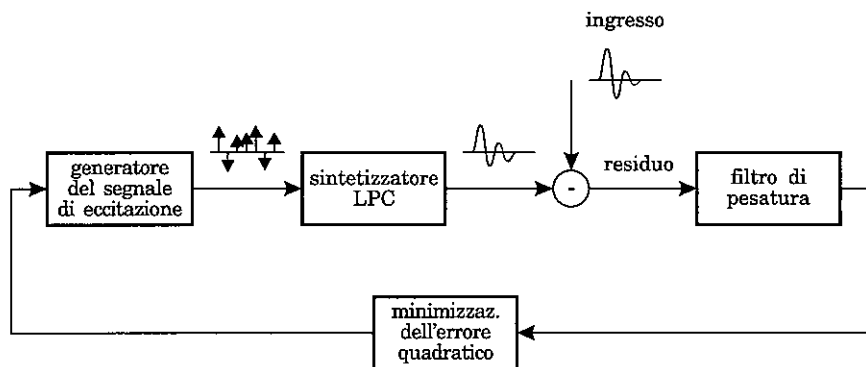



Fig. 6.6 - Codifica Multipulse.

Come al solito, il problema che è necessario affrontare è quello della codifica dell'eccitazione. Una possibile soluzione è quella di ricorrere ad un dizionario (codebook) di possibili frammenti di eccitazione scelti, ad esempio, in maniera casuale. Confrontando il segnale d'errore per ciascun blocco con quelli presenti nel dizionario, è possibile ricavare la sua migliore approssimazione, trasmettendone l'indice (Code-Excited Linear Prediction: CELP).

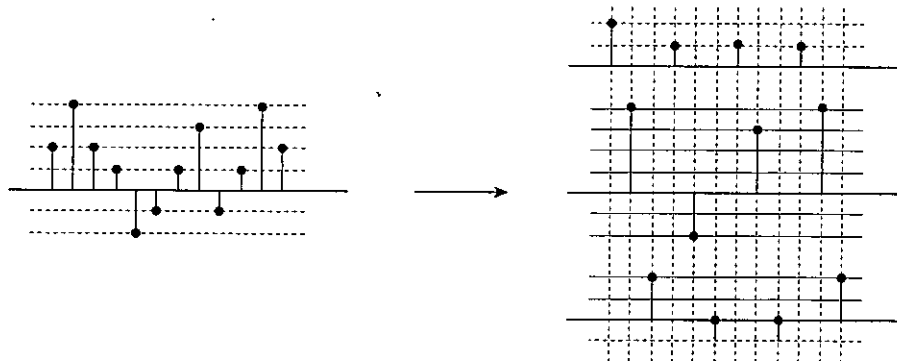



Fig. 6.7 - Decimazione dell'eccitazione per codifica RPE.

Una differente soluzione prevede la decimazione del segnale d'errore (fig. 6.7). Considerando nel blocco dei campioni il sottoinsieme dei "più rappresentativi", è possibile ridurre il numero di campioni trasmessi, lasciando al filtro AR il compito della loro interpolazione. La scelta ottima dei campioni è un compito estremamente gravoso dal punto di vista computazionale.

Per ridurre il numero delle combinazioni possibili (e quindi sia la complessità che il numero di informazioni da trasmettere) si definisce un passo costante di decimazione "n" detta "griglia" dell'eccitazione. Le sottosequenze possibili in tal modo sono n, ottenute prendendo campioni equispaziati con origine su una delle n possibili scelte (da cui il nome di Regular Pulse Excitation: RPE).

La sottosequenza adottata nella codifica è quella che massimizza la funzione di energia

$$E_M = \mathrm{Max}_m \, [\, x_m(i) \,]^2 \qquad (6.10)$$

e la sua codifica è data dalla posizione sulla griglia e da una codifica (es.: APCM) della sequenza di campioni.

6.3.2 Standard ETSI GSM 06.10

La codifica audio adottata nello standard del Radio Mobile Digitale Pan-Europeo del Groupe Special Mobil (GSM) dell'European Conference of Post and Telecommunications Administrations (CEPT) (raccomandazione CEPT / GSM 06.10) è un esempio di codifica per modelli RPE. Gli scopi di

tale standard sono quelli di ridurre il flusso numerico al di sotto dei 16 kbit/s con un ritardo di codifica inferiore agli 80 ms.

L'algoritmo utilizzato per la codifica è il Regular-Pulse Excitation / Long Term Prediction (RPE-LPT). Il segnale vocale è ottenuto tramite un treno di impulsi interpolato da un filtro tale da ricostruire l'inviluppo dello spettro a breve termine del segnale. L'ingresso è un flusso numerico derivante da un campionamento a 8 kHz con una codifica lineare su 13 bit. Le fasi della codifica sono (fig. 6.8):

- pre-elaborazione (per migliorare la precisione e la stabilità degli algoritmi);
- analisi LPC a breve termine (per la determinazione dell'inviluppo dello spettro del segnale);
- analisi a lungo termine (per la determinazione della struttura a grana fine dello spettro del segnale;
- decimazione RPE (per la codifica dell'eccitazione del modello digitale della sorgente).

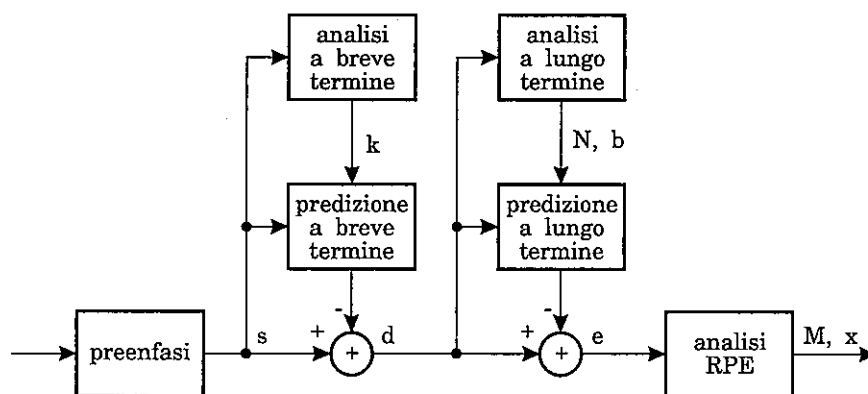



Fig. 6.8 - Diagramma del codificatore RPE-LTP.

La pre-elaborazione ha il fine di eliminare le componenti continue presenti nel segnale ed eseguire una pre-enfasi del segnale stesso, al fine di migliorare la stabilità degli algoritmi di filtraggio adattativo ed aumentare la precisione numerica (fig. 6.9).

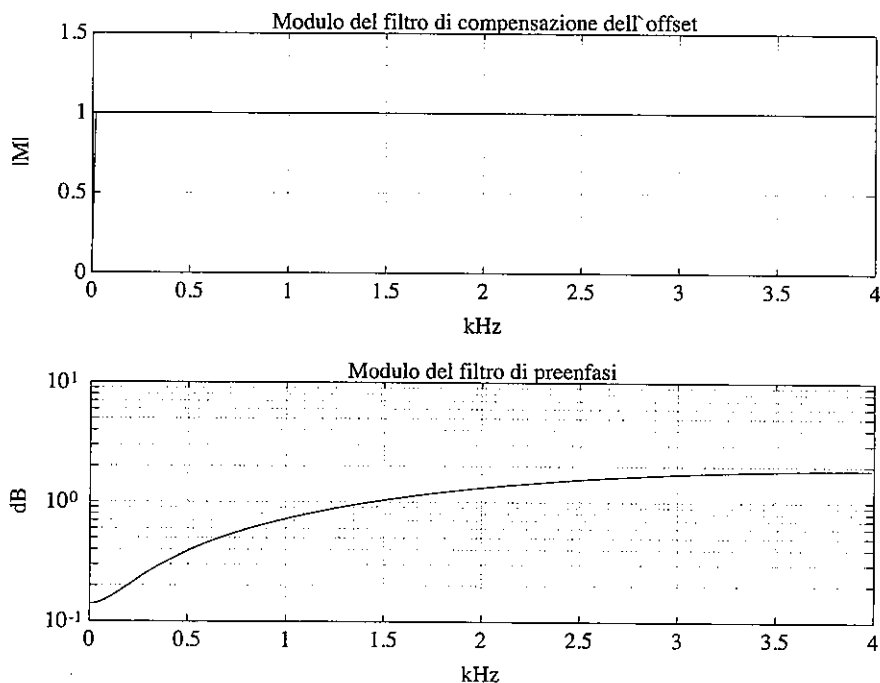



Fig. 6.9 - Modulo dei filtri ci compensazione dell'offset e di preenfasi.

Per l'analisi LPC il flusso numerico è segmentato in gruppi di 160 campioni (20 ms). Il modello della sorgente che si ricava è dell'ottavo ordine. I parametri del modello vengono ottenuti come coefficienti di riflessione di un modello FIR lattice tramite l'algoritmo di Schür [Appendice C]. La ricorsione viene forzatamente interrotta ad un passo precedente all'ottavo (restituendo di fatto un predittore di ordine inferiore ad 8) qualora si stiano producendo coefficienti maggiori in valore assoluto di 1; ciò è indice che l'algoritmo di stima sta divergendo e l'interruzione è richiesta al fine di garantire la stabilità del predittore.

Per la codifica, gli 8 coefficienti di riflessione ricavati sono poi convertiti in Log Area Ratios (LARs), meno sensibili agli errori di quantizzazione, definiti come

$$r_i = \log \frac{1 - k_i}{1 + k_i} \tag{6.11}$$

Essi sono legati ad un modello meccanico del tratto vocale tramite una sequenza di tubi senza perdite e debbono il loro nome al fatto di risultare legati al rapporto tra le aree di sezioni adiacenti [Rab78]

$$r_i = \log \frac{A_{i+1}}{A_i} \tag{6.12}$$

I LARs sono quantizzati linearmente, ma, a causa della differente dinamica, la loro codifica avviene con un numero variabile di bit (6 per i primi 2 e scalando fino a 3 per gli ultimi due).

Per l'analisi a breve termine, i LARs sono decodificati, riottenendo i coefficienti del modello AR (affetti da rumore di quantizzazione). Vengono, quindi, calcolati 160 campioni della funzione d'errore tra ingresso e segnale stimato. Per evitare transizioni spurie nel passaggio tra due segmenti di segnale, i coefficienti utilizzati per i primi 40 campioni del blocco sono ottenuti interpolando linearmente i nuovi LAR con i precedenti. In particolare, al passo i-esimo per i campioni n-esimi, i coefficienti utilizzati sono ottenuti come

$$\begin{cases} 0.75\,LAR^{(i-1)} + 0.25\,LAR^{(i)}\,; & n \in [0..12] \\ 0.50\,LAR^{(i-1)} + 0.50\,LAR^{(i)}\,; & n \in [13..26] \\ 0.25\,LAR^{(i-1)} + 0.75\,LAR^{(i)}\,; & n \in [27..39] \end{cases} \tag{6.13}$$

Il residuo di predizione a breve termine è sottoposto ad analisi a lungo termine (fig. 6.10). Per ogni intervallo di 5 ms (40 campioni) si calcola il periodo N (codificato su 7 bit), tramite il massimo della funzione di autocorrelazione

$$R(\lambda) = \sum_{i=0}^{39} d(k_j + i)\, d(k_j + i - \lambda)$$

$$R(N) = \mathrm{Max}[\, R(\lambda)\,];\ \lambda = 40\ ..\ 120 \tag{6.14}$$

viene inoltre calcolato il guadagno a lungo termine β (codificato su 2 bit), tramite il rapporto tra ampiezza e valore efficace

$$\beta = \frac{R(N)}{\sum_{i=0}^{39} d'^2(k_j + i - N)} \tag{6.15}$$

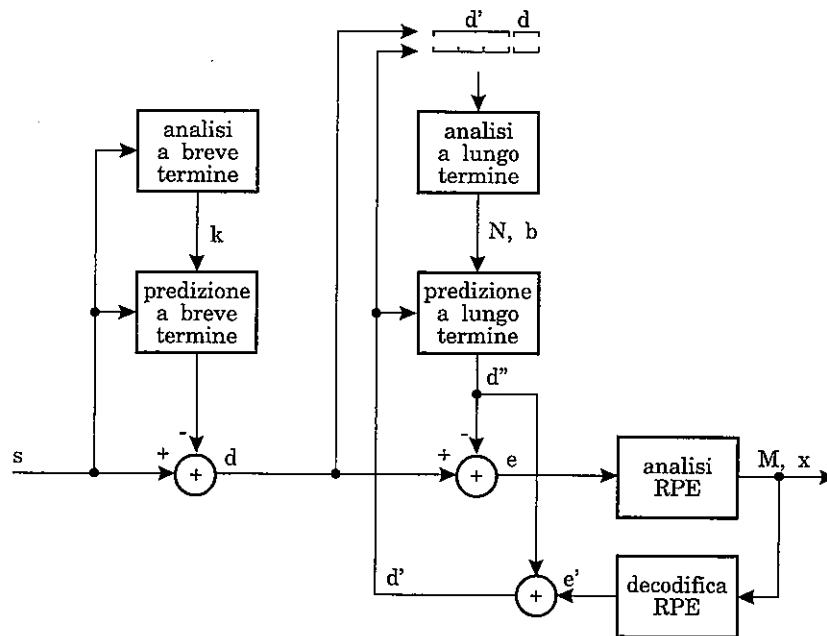



Fig. 6.10 - Sezione APC del codificatore GSM.

La predizione ottenuta dall'analisi a lungo termine viene sottratta all'errore ottenuto dall'analisi a breve termine. Il residuo viene filtrato tramite un FIR dell'11° ordine (sempre per sagomare lo spettro dell'errore) ed il risultato sottoposto a codifica RPE. La codifica avviene decimando il residuo a lungo termine per ciascun sottoblocco, scartando 2 campioni su 3 e scegliendo, tra le 4 possibili sequenze di 13 campioni sfasate di un campione ciascuna, quella che massimizza la funzione di energia

$$E_M = \mathrm{Max}_m \, [\, x_m(i) \,]^2 \quad m = 0, 1, 2, 3 \tag{6.16}$$

Si fa notare come la prima e l'ultima sequenza risultino coincidenti a meno di uno sfasamento di un campione. Tali campioni sono codificati tramite una quantizzazione adattativa semplificata, ottenuta normalizzando ciascun campione rispetto al massimo della sequenza (codificato su 6 bit) ed esprimendone l'ampiezza tramite una quantizzazione non uniforme su 3 bit.

Riepilogando, per ogni blocco di 160 campioni su 13 bit (2080 bit) vengono prodotte le grandezze riportate in tabella 6.2.

Algoritmo	Parametro	Bit per sottosegmento (40 campioni)	Bit per segmento (160 campioni)
pred. breve termine	LAR		36
pred. lungo termine	β	2	8
pred. lungo termine	N	7	28
RPE (griglia)	M	2	8
RPE (eccitazione)	x_{max}	6	24
RPE (eccitazione)	x_m	3*13	156
	Totale	56	260

Tab. 6.2 - Componenti del flusso GSM.

Il rapporto di compressione che ne deriva è di 8:1. Rispetto al log-PCM il flusso prodotto (50 segmenti al secondo, cioè 13 kbit/s) ha un rapporto di compressione di 5:1.

6.4 CODIFICA MULTIPULSE

La tecnica di codifica Multipulse rappresenta la tecnica di Analisi per Sintesi introdotta storicamente per prima in letteratura. La sua formulazione si deve ad Atal e Remde nel 1982 [Ata82]. La tecnica può essere interpretata come uno sviluppo della tecnica APC (Adaptive Predictive Coding) di cui si fornisce nel seguito un breve cenno.

L'algoritmo APC [Ata79] è simile all'algoritmo ADPCM da cui si differenzia per la presenza di un predittore a lungo termine ed un filtro di sagomatura spettrale del rumore di quantizzazione (Noise Shaping). In alcuni casi i coefficienti di predizione a breve termine sono calcolati sul frame corrente, codificati e trasmessi. Di questa tecnica sono state presentate alcune varianti come quella proposta da Itakura [Hon84] in cui si introduce una allocazione dinamica dei bit riservati alla quantizzazione, o quella proposta in [Mak79] in cui il predittore a lungo termine è rimosso. Atal ha poi proposto una ulteriore modifica allo schema introducendo la procedura di Center Clip (CC) del segnale residuo di predizione, giungendo allo schema illustrato in figura 6.11 [Ata80] in cui il blocco NS è responsabile della sagomatura spettrale del rumore di quantizzazione, mentre Ps e Pl rappresentano i predittori a breve e lungo termine rispettivamente.

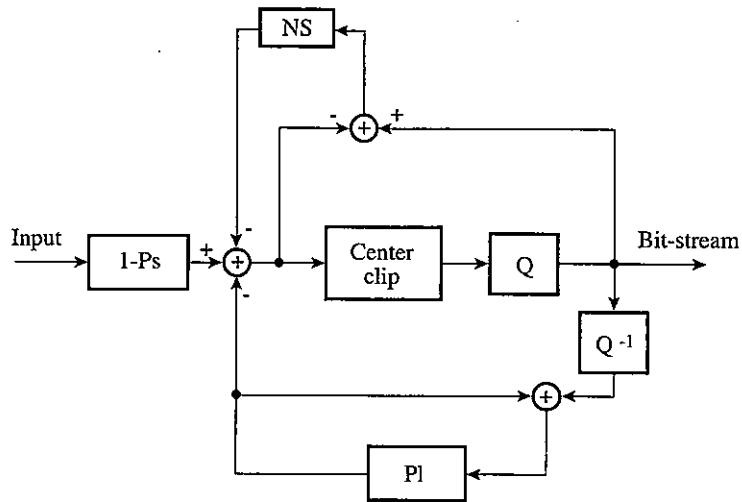



Fig. 6.11 - Schema del codec APC-NS con Center Clip.

La procedura di Center Clip consiste nel porre a zero tutti i campioni con ampiezza inferiore ad una certa soglia (Th) e quindi corrisponde ad introdurre un blocco con funzione di trasferimento non lineare illustrata in figura 6.12.

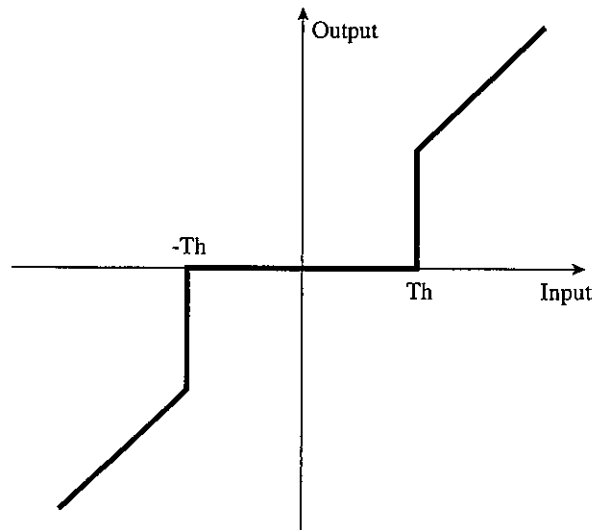



Fig. 6.12 - Funzione di trasferimento del blocco Center Clip.

Mentre il blocco di sagomatura del rumore di quantizzazione può essere facilmente interpretato come un primo passo per introdurre la funzione di pesatura percettiva, che è un elemento caratteristico degli schemi ABS, il blocco di CC corrisponde a creare un segnale di eccitazione in cui solo alcuni campioni sono non-nulli, che è esattamente il tipo di segnale di eccitazione usato nello schema Multipulse.

Lo schema a blocchi di principio del codec Multipulse è riportato in figura 6.13

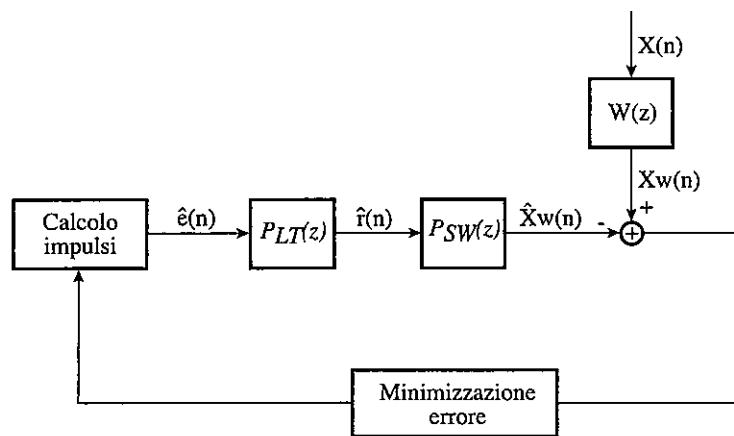



Fig. 6.13 - Schema a blocchi di un codec Multipulse.

La struttura è la stessa del codec CELP, con la differenza che la procedura di minimizzazione dell'errore è utilizzata, in questo caso, come criterio per determinare la posizione e la ampiezza di un certo numero di impulsi, anziché determinare la parola ottima del codebook.

Le formule utilizzate per il calcolo dei vari impulsi sono le stesse impiegate per lo schema CELP. La procedura consiste nel calcolare sequenzialmente posizione ed ampiezza degli impulsi, considerando ad ogni iterazione l'effetto degli impulsi calcolati precedentemente. La procedura è quindi meno complessa di quella dello schema CELP, in quanto il numero di calcoli di distorsione risulta pari al numero di impulsi da introdurre, che può essere dell'ordine di 5 impulsi su 40 campioni. Inoltre, questo metodo consente facilmente di tenere in conto degli effetti della quantizzazione, in quanto ad ogni iterazione i nuovi impulsi possono essere quantizzati. Tuttavia, al crescere del numero di impulsi, la procedura diventa meno efficiente e diventa

essenziale considerare congiuntamente l'effetto dei diversi impulsi. È intuitivo che una procedura completamente esaustiva risulta analoga a quella impiegata nello schema CELP in cui il vocabolario è di tipo sparso. Una soluzione intermedia in termini di complessità consiste nel riottimizzare le ampiezze dei vari impulsi dopo avere determinato sequenzialmente le loro posizioni.

Anche per lo schema Multipulse, come per l'RPE ed il CELP, l'introduzione di un filtro di predizione a lungo termine comporta un significativo miglioramento delle prestazioni.

A fronte di un discreto successo iniziale, la tecnica Multipulse è stata successivamente soppiantata dalla tecnica CELP che, viceversa, è stata impiegata, come si vedrà, in numerosi standard di codifica.

6.5 CODIFICA CELP

La tecnica di codifica CELP è sicuramente la tecnica che ha avuto più successo nell'ultimo decennio ed è stata studiata da quasi tutti i laboratori di ricerca sulla codifica della voce del mondo. A riprova del suo successo, diversi standard di codifica della voce sono basati su questo algoritmo. La tabella 6.3 riporta i principali standard di codifica basati appunto sulla tecnica CELP.

Sigla	Ente	Descrizione
G.728	ITU-T	LD-CELP - Codifica CELP Low-Delay a 16 kbit/s
G.729	ITU-T	CELP a 8 kbit/s
G.723	ITU-T	CELP a 5.27 e 6.3 kbit/s per codifica dell'audio in multimedia
IS-54	ANSI	CELP a 7.95 kbit/s. Full-rate per sistemi cellulari in Nord-America (basati su D-AMPS)
IS-96	ANSI	CELP a velocità variabile 1.2 - 9.6 kbit/s. Codifica per sistemi cellulari in Nord-America (basati su CDMA)
US-1	ANSI/ETSI	CELP a 12.2 kbit/s. Codifica Enhanced Full-Rate per GSM
GSM-HR	ETSI	CELP a 5.6 kbit/s. Half-rate per il sistema cellulare GSM
FS 1016	NATO (DoD)	CELP a 4.8 kbit/s

Tab. 6.3 - Elenco dei principali standard di codifica basati sulla tecnica CELP.

La tecnica è stata introdotta in letteratura nel 1984 come una particolare tecnica di codifica stocastica [Ata84] e come algoritmo di pattern classification [Cop84]. Il termine CELP si deve ad una pubblicazione di Atal del 1985 [Sch85]. Dalle prime implementazioni della tecnica con complessità altissime si è pervenuti in alcuni anni a soluzioni sempre più semplificate, fino alle implementazioni attuali con complessità che si aggirano attorno ai 20 MOPS e quindi realizzabili utilizzando un solo DSP.

6.5.1 L'algoritmo CELP

Lo schema base dell'algoritmo CELP (fig. 6.14) è costituito da una catena di sintesi inserita in un anello di controreazione che è utilizzato per determinare il segnale di eccitazione $\hat{e}(n)$ più consono a rappresentare il segnale ricostruito $\hat{X}(n)$. La catena di sintesi costituisce lo schema del ricevitore, una cui copia locale è inserita all'interno del trasmettitore. Per semplicità in questo schema di principio non sono riportati i blocchi relativi al calcolo dei parametri dei filtri ed alla loro quantizzazione, ma essi costituiscono parte integrante del trasmettitore.

Il sistema è costituito da cinque blocchi principali, riportati in figura 6.14, che sono illustrati in dettaglio nel seguito.

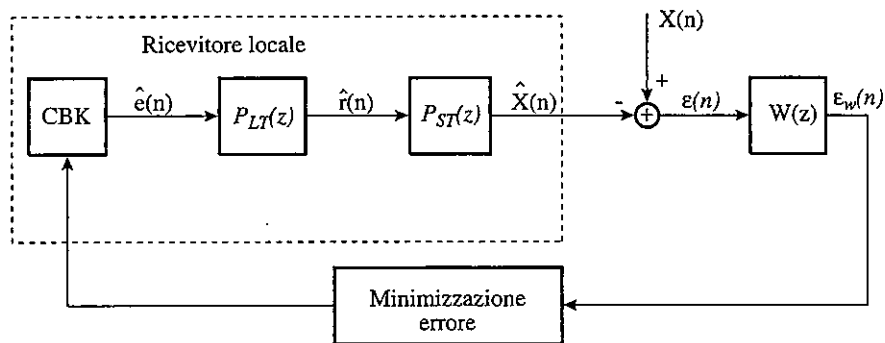



Fig. 6.14 - Schema di principio dell'algoritmo CELP.

Il primo blocco è una tabella, o codebook (CBK), contenente una collezione di vettori di lunghezza opportuna, tali da essere impiegati per

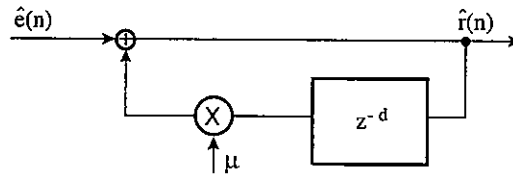



Fig. 6.15 - Schema a blocchi del filtro di sintesi a lungo termine.

realizzare il segnale di eccitazione $\hat{e}(n)$. La tabella è simile a quelle impiegate per la quantizzazione vettoriale e la procedura di selezione del vettore può essere vista come un processo di quantizzazione vettoriale, in cui il criterio da minimizzare (misura di distorsione) è legato ad un segnale trasformato. La lunghezza dei vettori varia da schema a schema, da un minimo di 4 campioni ad un massimo di 60. La dimensione del vocabolario, una volta fissata la lunghezza dei vettori, determina il contributo alla velocità di trasmissione necessario per la rappresentazione del segnale di eccitazione. Solitamente tale dimensione non supera i 10 bit (1024 parole). Per dimensioni superiori, in genere sono state utilizzate strutture multi-stadio. Relativamente alla generazione di tale tabella, sono state proposte numerose soluzioni e se ne parlerà più in dettaglio in seguito.

Il segnale di eccitazione, estratto dalla tabella, alimenta un filtro lineare di sintesi $P_{LT}(z)$ che tiene conto della correlazione a lungo termine del segnale vocale. Nella sua realizzazione più semplice, tale filtro ha la struttura riportata in figura 6.15 e quindi funzione di trasferimento data da

$$P_{LT}(z) = \frac{1}{1 - \mu \cdot z^{-d}} \tag{6.17}$$

in cui il parametro d tiene conto della periodicità a lungo termine del segnale vocale ed il coefficiente μ del grado di correlazione. Il parametro d ha il significato fisico di periodo fondamentale del segnale vocale (pitch) e pertanto il filtro di correlazione a lungo termine è responsabile dell'inclusione nel segnale $\hat{r}(n)$ dell'informazione legata alla vibrazione delle corde vocali.

Essendo legato al periodo fondamentale, il ritardo d assume valori che sono confinati, nella grande maggioranza dei casi, nell'intervallo tra 20 e 160 campioni. Questi limiti sono relativi, alla frequenza di campionamento di 8 kHz,

a frequenze del pitch comprese tra 50 e 400 Hz.I parametri del filtro di sintesi a lungo termine possono essere calcolati in catena aperta sul segnale vocale con i metodi classici [Hes83], oppure possono essere determinati in catena chiusa, come per il segnale di eccitazione, minimizzando anche in questo caso l'errore pesato percettivamente. In ogni caso devono essere quantizzati e trasmessi al ricevitore ad ogni frame o anche più rapidamente, in considerazione del cambiamento delle caratteristiche del segnale. Si vedrà in seguito come anche questo blocco possa essere sostituito da strutture più sofisticate, al fine di consentire una miglior fedeltà di riproduzione della frequenza fondamentale.

Il segnale $\hat{r}(n)$ così prodotto prende il nome di segnale residuo ricostruito ed alimenta un secondo filtro lineare di sintesi, responsabile dell'introduzione dell'informazione legata alle formanti e cioè relativa alla correlazione a breve termine. Di tale filtro e dei dettagli relativi al calcolo e trasformazione dei parametri si è già parlato nei capitoli precedenti. In alcuni casi particolari, con lunghezza del vettore molto corta (4 o 5 campioni), i parametri possono essere calcolati sul segnale vocale ricostruito (in backward) e pertanto non devono essere trasmessi in quanto possono essere ricavati anche al ricevitore. È il caso ad esempio del codec ITU-T G.728 (LD-CELP). L'uscita del filtro $P_{LT}(z)$ rappresenta il segnale vocale ricostruito $\hat{x}(n)$.

Il blocco successivo è costituito dal filtro lineare tempo variante W(z), impiegato per sagomare spettralmente il segnale errore di quantizzazione $\varepsilon(n)$. L'idea dell'introduzione di questo blocco trae origine dall'osservazione che a basse velocità di trasmissione, il rapporto segnale su rumore ottenibile è di soli pochi dB. In queste condizioni, un errore di quantizzazione a spettro piatto, quale si otterrebbe senza nessuna sagomatura del segnale errore $\varepsilon(n)$, porterebbe alla situazione illustrata in figura 6.16 in cui l'SNR può essere molto alto in corrispondenza della prima formante, ma anche negativo in certe porzioni dello spettro.

L'idea è quindi quella di sagomare spettralmente il segnale $\varepsilon(n)$, al fine di ottenere un segnale errore del quale minimizzare l'energia nel processo di selezione del vettore di eccitazione. La funzione di trasferimento impiegata nel blocco W(z) è quella originariamente proposta da Atal negli studi sulla tecnica di codifica APC (Adaptive Predictive Coding) [Ata79] e successivamente impiegata nel codificatore Multipulse. Il blocco prende il nome di filtro di pesatura spettrale ed è definito dalla funzione di trasferimento

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}} \tag{6.18}$$

Questa funzione ha lo svantaggio di non tenere sufficientemente in conto i fenomeni di mascheramento spettrale operati dall'apparato uditivo, tuttavia consente una semplice adattatività dei parametri alle caratteristiche del segnale vocale ed ha il grande vantaggio di consentire una significativa semplificazione dello schema di codifica, come sarà chiaro nel seguito.

I coefficienti ai sono quelli dell'analisi lineare a breve termine ed il coefficiente $\gamma \in [0, 1]$ ha l'effetto di allargare la banda delle formanti di un fattore $\Delta f = -\left(f_s/\pi\right) \cdot (\ln\gamma)$, senza modificarne la posizione.

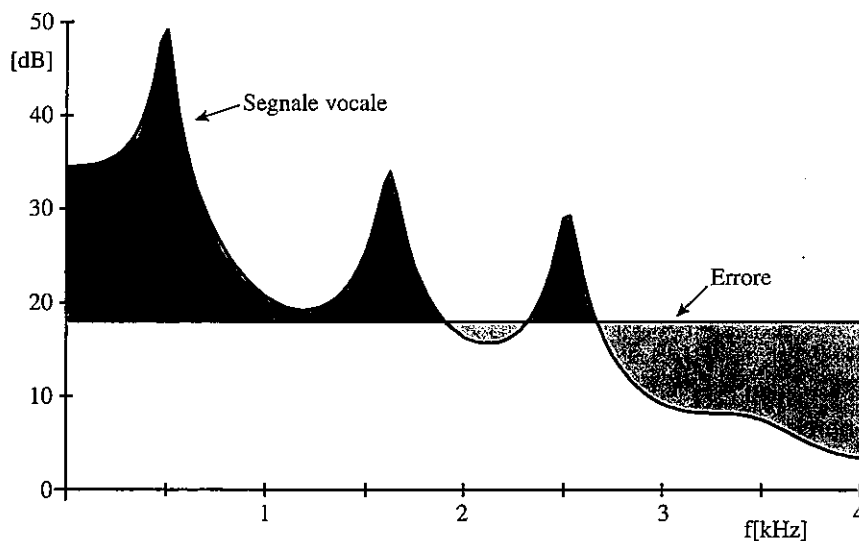



Fig. 6.16 - Spettro del segnale vocale e del segnale errore di quantizzazione nel caso di assenza di sagomatura spettrale.

L'effetto di pesatura del segnale errore risulta tanto più pronunciato tanto più γ è prossimo a 0, nel qual caso W(z) collassa nella funzione di trasferimento del filtro di analisi LPC. A titolo esemplificativo, la figura 6.17

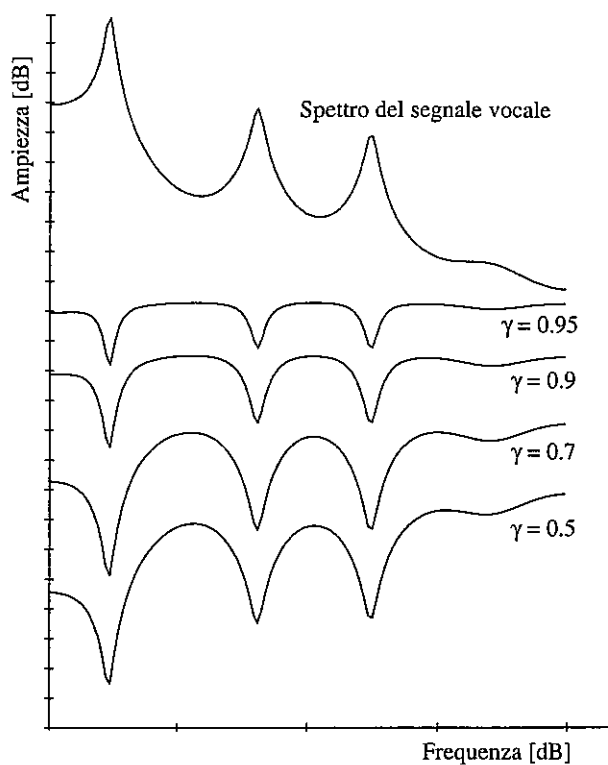



Fig. 6.17 - Andamento del modulo del filtro di pesatura spettrale per diversi valori del parametro γ.

riporta l'andamento del modulo di W(z) in funzione della frequenza, per diversi valori di γ. I valori tipicamente più utilizzati sono tra 0.8 e 0.9.

L'ultimo blocco dell'algoritmo di codifica CELP è costituito dal calcolo dell'energia del segnale errore pesato. Questa è calcolata per ognuno dei vettori di eccitazione raccolti nella tabella, in modo da selezionare quello corrispondente all'energia minima. Ovviamente considerando che nella catena ci sono dei filtri con memoria e con coefficienti tempo varianti, bisognerà opportunamente fissare le condizioni iniziali di ognuno di essi in modo da considerare, nella ricerca, ogni vettore nelle stesse condizioni.

Sebbene la struttura presentata sia utile per introdurre teoricamente l'algoritmo, essa non si presta ad essere implementata direttamente, in quanto richiederebbe una complessità di calcolo molto elevata. Sono state quindi ideate soluzioni alternative che consentono significative riduzioni di complessità

6.5.2 Tecniche di riduzione di complessità

La considerazione alla base della modifica dello schema teorico è che tutte le trasformazioni operate sul vettore eccitazione devono essere fatte, in fase di ricerca del minimo, per ogni vettore del codebook, pertanto ogni riduzione di tali operazioni determina una riduzione di complessità notevole.

La prima operazione consiste nel trasferire il blocco W(z) alla sinistra del sommatore in figura 6.14 e quindi riportandolo sui due rami. A questo punto si osserva che il numeratore della funzione W(z) è uguale alla funzione di trasferimento di $P_{ST}(z)$, e pertanto semplificabile. Si giunge quindi allo schema di figura 6.18 in cui il blocco $P_{SW}(z)$ ha funzione di trasferimento

$$P_{SW}(z) = \frac{1}{A(z/\gamma)} = \frac{1}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}} \qquad (6.19)$$

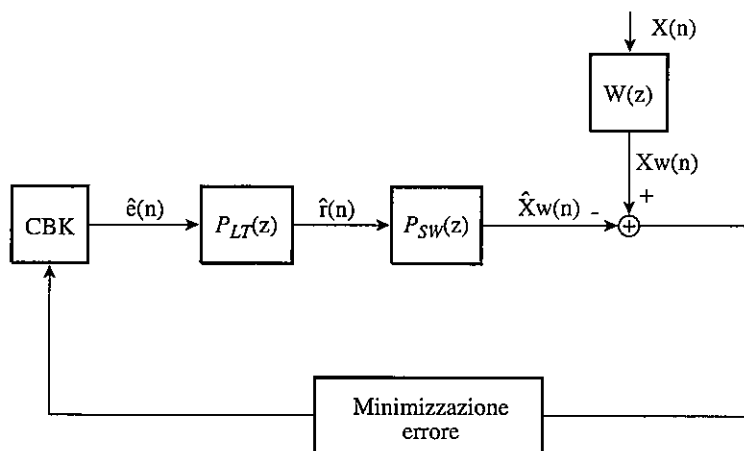



Fig. 6.18 - Semplificazione dello schema CELP.

Sfruttando la sovrapposizione degli effetti, la catena dei due filtri di sintesi può essere sdoppiata in due componenti: una responsabile del contributo delle memorie e con ingresso nullo ed un'altra responsabile del contributo del vettore del codebook, ma senza memorie dei filtri. Questa operazione porta allo schema di figura 6.19.

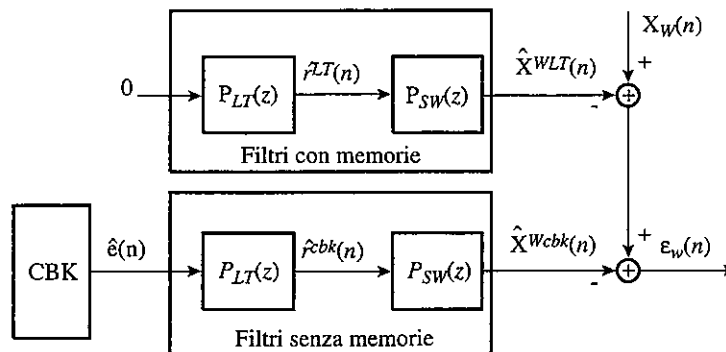



Fig. 6.19 - Schema CELP semplificato con separazione dei contributi dovuti ai filtri di sintesi.

Ricordando ora che la funzione di trasferimento del filtro di sintesi a lungo termine è tale da considerare solo quei campioni con ritardo maggiore di d, ne consegue che, qualora la lunghezza dei vettori del codebook sia minore di d minimo, il blocco $P_{LT}(z)$ non fornisce alcun contributo, dato che è senza memorie e quindi $\hat{r}_{cbk}(n) = \hat{e}(n)$. Il segnale $\hat{r}_{LT}(n)$, uscita del filtro di sintesi a lungo termine e con ingresso nullo, assume il significato di contributo del pitch al segnale di eccitazione e viene spesso identificato come contributo dovuto ad un codebook adattativo. Infine, se si ripete l'operazione di sdoppiamento anche considerando questo segnale $\hat{r}_{LT}(n)$, si perviene allo schema finale effettivamente utilizzato in pratica, che è riportato in figura 6.20.

In riferimento allo schema iniziale di figura 6.18, valgono le relazioni

$$\begin{aligned} \hat{X}_W(n) &= \hat{X}_{W0}(n) + \hat{X}_{WLT}(n) + \hat{X}_{Wcbk}(n) \\ \hat{r}(n) &= \hat{r}_{LT}(n) + \hat{r}_{cbk}(n) \end{aligned} \tag{6.20}$$

e nel caso in cui $d \geq L_s$,

$$\hat{r}(n) = \hat{e}(n) + \hat{r}_{LT}(n) \Leftrightarrow d \geq L_s \tag{6.21}$$

con Ls lunghezza del vettore del codebook.

Come già detto questa struttura si presta ad implementazioni di gran lunga più efficienti della struttura base. Infatti in questo caso l'operazione di ricerca della parola del codebook ottima comprende essenzialmente una

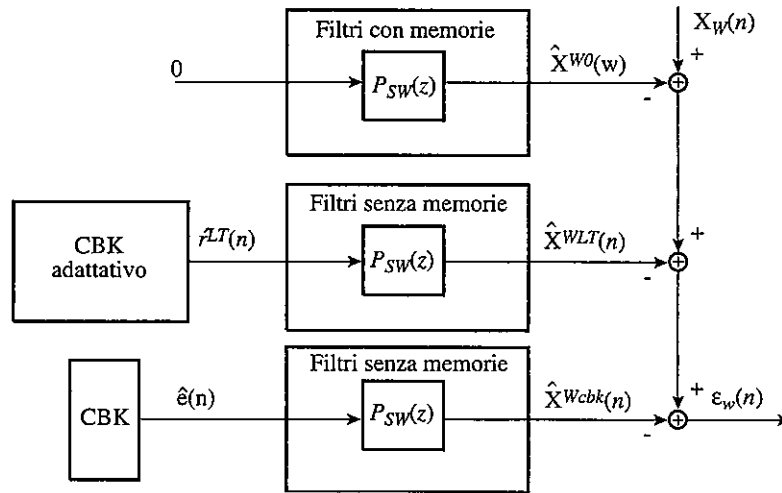



Fig. 6.20 - Schema CELP semplificato.

operazione di filtraggio senza memoria che può essere sostituita dal prodotto del vettore del codebook per la matrice caratteristica del filtro. Tutte le altre operazioni sono indipendenti dai vettori della tabella e pertanto possono essere effettuate una sola volta per tutti i vettori da analizzare.

Supponendo di calcolare i parametri del codebook adattativo sul segnale vocale (open loop), il calcolo dell'eccitazione ottima consiste nella minimizzazione dell'energia del segnale errore pesato $\varepsilon_W(n)$ e cioè

$$\min_j(E(j)) = \min_j \| \varepsilon_W(j) \|^2 = \min_j \sum_{n=1}^{L_s} (\varepsilon_W(j,n))^2 \qquad (6.22)$$

in cui j rappresenta l'indice della generica parola del codebook ed Ls è la lunghezza del vettore. Questa viene solitamente indicata con il termine di *subframe* in quanto sottomultiplo della durata del *frame* che costituisce l'intervallo di tempo sul quale sono calcolati i parametri del filtro di sintesi LPC.

L'energia dell'errore di quantizzazione pesato è calcolata con la relazione

$$E(j) = \sum_{n=1}^{L_s} \left(X_W(n) - \hat{X}_{W0}(n) - \hat{X}_{wLT}(n) - \hat{X}_{wcbk}(j,n) \right)^2 \qquad (6.23)$$

in cui i primi tre termini non dipendono dalla specifica parola del codebook e quindi possono essere calcolati una sola volta off-line dalla procedura di ricerca. Tralasciando la dipendenza dell'errore dalla generica parola j, e ponendo $X_{Woff} = X_W(n) - \hat{X}_{W0}(n) - \hat{X}_{wLT}(n)$, si ottiene

$$E = \sum_{n=1}^{L_s} \left(X_{Woff}(n) - \hat{X}_{wcbk}(j,n) \right)^2$$
$$= \sum_{n=1}^{L_s} \left(X_{Woff}(n) - g_{cbk} \cdot I(n) \right)^2 \qquad (6.24)$$

dove I(n) rappresenta il generico vettore del codebook filtrato con il filtro di sintesi a breve termine senza memorie. Derivando rispetto a g_{cbk} e ponendo la derivata uguale a zero, si ottiene la relazione del fattore di scala ottimo per il generico vettore, e cioè

$$g_{cbk} = \frac{\sum_{n=1}^{L_s} X_{Woff}(n) \cdot I(n)}{\sum_{n=1}^{L_s} (I(n))^2} \qquad (6.25)$$

Sostituendo la relazione del fattore di scala ottimo in quella dell'energia, si ottiene l'espressione dell'energia ottenibile con il fattore di scala ottimo

$$E = \sum_{n=1}^{L_s} (X_{Woff}(n))^2 - \frac{\left(\sum_{n=1}^{L_s} X_{Woff}(n) \cdot I(n) \right)^2}{\sum_{n=1}^{L_s} (I(n))^2} \qquad (6.26)$$

Dato che il primo termine è costante per ogni vettore del codebook, il processo di ricerca del vettore ad energia minima consiste quindi nella massimizzazione del secondo termine. Pertanto il vettore di eccitazione ottimo è quello che massimizza il termine

$$\max_{j} \left\{ \frac{\left(\sum_{n=1}^{L_s} X_{Woff}(n) \cdot I(j,n) \right)^2}{\sum_{n=1}^{L_s} (I(j,n))^2} \right\} \qquad (6.27)$$

Sulla base di questa relazione, sono state poi introdotte in letteratura ulteriori semplificazioni che comprendono, ad esempio, la memorizzazione del denominatore per ogni vettore e soprattutto il confronto tra rapporti, il quale consente di evitare il calcolo esplicito della divisione per ogni vettore.

Resta da osservare che se si vuole tenere in conto degli effetti della quantizzazione del fattore di scala nella ricerca del vettore minimo, non si può utilizzare l'ultima relazione, ma si deve esplicitamente inserire il valore del fattore di scala, con conseguente aumento della complessità di calcolo.

Queste relazioni sono valide per il calcolo del vettore di eccitazione nell'ipotesi i parametri del filtro del pitch siano disponibili e costanti per ogni vettore. Un miglioramento delle prestazioni è ottenibile calcolando anche i parametri di tale filtro con la stessa procedura (closed loop). Il metodo è simile e consiste nella minimizzazione dell'energia del segnale XWoff(n). In questo caso l'incognita sarà il ritardo d e si può vedere facilmente che si ottengono espressioni analoghe.

Seppur questa procedura di minimizzazione sequenziale consenta un miglioramento delle prestazioni, la soluzione ottima consiste ovviamente nella determinazione congiunta dei contributi dei due rami minimizzando l'energia del segnale $\varepsilon_W(n)$. Quest'ultima possibilità può essere ottenuta con la procedura di ortogonalizzazione delle componenti che sarà descritta nel seguito.

6.5.3 Varianti allo schema CELP

A seguito del successo della tecnica CELP, sono state proposte in letteratura numerose varianti, alcune tese a ridurre la complessità di calcolo ed altre invece tese a migliorare le prestazioni.

Tipi di codebook

Nelle prime realizzazioni dello schema CELP, il codebook era costituito da vettori estratti da una sequenza di campioni distribuiti con d.d.p. alla Gauss [Ata84], oppure realizzato impiegando una versione modificata dell'algoritmo LBG per la generazione di quantizzatori vettoriali [Cop84]. La motivazione per l'utilizzo di una sequenza Gaussiana discende dalla considerazione che l'errore residuo derivante dalle predizioni a breve e lungo termine ha una distribuzione statistica di questo tipo.

Successivamente sono stati introdotte strutture di codebook che consentono semplificazioni di calcolo e di capacità di memoria necessaria. È il caso ad esempio di codebook cosiddetti ciclici o overlappati [Lin88] in cui le varie parole differiscono per uno o pochi campioni, oppure codebook sparsi binari o ternari [Dav86] in cui solo alcuni campioni del vettore hanno ampiezza non nulla. Tali codebook sono ottenuti con una procedura di center-clip e ponendo ad ampiezza unitaria, con segno, i campioni non nulli.

Una menzione particolare spetta ai vocabolari definiti a chiavi [Cel89] o a codici algebrici [Ado87] per i quali non è necessaria la memorizzazione esplicita delle parole in quanto possono essere ottenute con shift opportuni di alcuni impulsi base. Infine nel codec VSELP (Vector Sum Excited Linear Predictive) [Ger90] il codebook è costituito da vettori ortogonali e linearmente indipendenti.

Self-excited

La tecnica di codifica SEV (Self Excited Vocoder) [Ros86] si differenzia dallo schema CELP in quanto il contributo di eccitazione del codebook in questo caso non esiste. Vale a dire che il segnale di eccitazione è formato utilizzando la sola informazione legata alla correlazione a lungo termine. In questo caso quindi è necessaria una opportuna inizializzazione delle memorie, altrimenti il segnale in uscita sarebbe sempre nullo. La tecnica consente velocità di trasmissione molto basse e possono essere impiegati più di un filtro di sintesi a lungo termine, determinando quindi più contributi che possono essere trattati sequenzialmente o congiuntamente. È evidente che ulteriori predittori oltre al primo, perdono il significato fisico di essere legati alla frequenza di vibrazione delle corde vocali e costituiscono piuttosto un modo alternativo per generare la sequenza di eccitazione.

Isolatamente tale tecnica ha avuto scarso successo, mentre acquista significato in tecniche di codifica ad oggetti, in cui si presta ad essere impiegata per quei tratti di segnale fortemente vocalizzati.

Pitch frazionario

L'andamento nel tempo della frequenza fondamentale del parlatore, o barra sonora in termini fonologici, costituisce l'informazione prosodica della comunicazione ed è pertanto responsabile dell'intonazione del messaggio. Tale andamento (pitch contour) è smussato nel tempo e le variazioni del pitch sono graduali e continue, al limite campione per campione. L'ipotesi di considerare tale valore costante per un frame o anche per un singolo subframe è chiaramente una approssimazione.

In particolare tale approssimazione comporta nello schema CELP dei possibili problemi che non consentono di sfruttare pienamente le potenzialità dello schema. Un esempio dei possibili problemi nel considerare il pitch costante è schematizzato in figura 6.21.

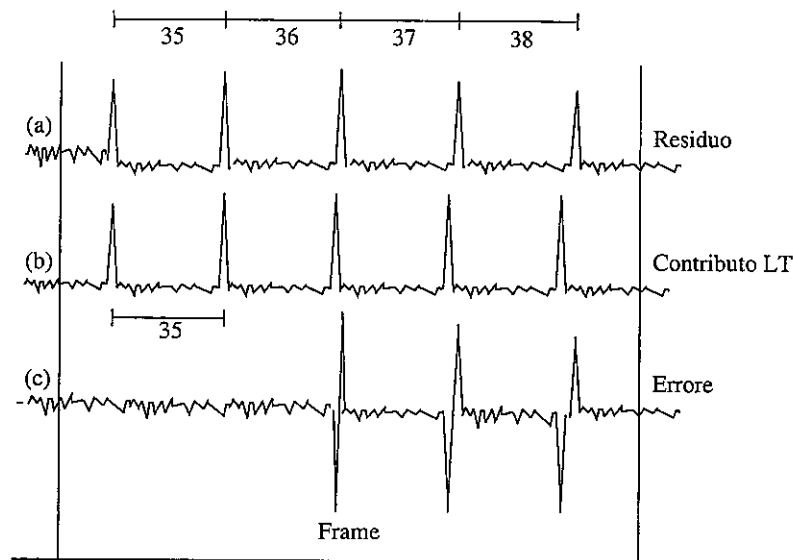



Fig. 6.21 - Andamento nel tempo del segnale residuo (a), la sua versione ricostruita dovuta al contributo LT(b) e l'errore tra i due segnali (c).

La figura riporta l'esempio di un possibile frame con valore del periodo di pitch stimato di 35 campioni. In realtà i quattro periodi nel frame hanno valori che vanno da 35 a 38 variando gradualmente. Utilizzando uno schema CELP, il contributo dato del predittore a lungo termine con ritardo d=35 sarà del tipo del segnale illustrato in figura 8b in cui la periodicità è fissata a 35 campioni. In questo caso il segnale errore XWoff(n)sarà del tipo in figura 8c e cioè presenterà una forte ampiezza in corrispondenza del disallineamento dei picchi dovuto alla non perfetta rappresentazione del pitch contour. In queste condizioni, il contributo del codebook sarà impiegato principalmente per sopperire a tale malfunzionamento e non già a recuperare l'errore impredicibile. Sebbene questo esempio sia estremo, in quanto le variazioni del pitch sono solitamente più limitate, lo stesso problema, in forma meno evidente, occorre nel considerare un valore di periodo intero senza tenere in conto valori frazionari mascherati dalla frequenza di campionamento.

Una prima soluzione semplice al problema è quella di considerare predittori a lungo termine con più di un coefficiente [Omo88]. Una soluzione più sofisticata consiste nel considerare valori del pitch frazionari ottenuti con un sovraccampionamento locale dei segnali in questione. La procedura solitamente impiegata in questi casi consiste nell'effettuare una prima stima del valore del pitch intero, mantenendo eventualmente più di un candidato, e quindi affinare la ricerca per valori frazionari nell'intervallo dei valori preselezionati.

Ortogonalizzazione

La procedura di ortogonalizzazione è impiegata per calcolare i contributi long-term e del codebook, minimizzando l'energia dell'errore pesato $\varepsilon_W(n)$. Più precisamente la procedura consiste nel modificare opportunamente il contributo del codebook in modo tale che la determinazione sequenziale dei due fattori di scala gckb e gLT sia equivalente alla loro ottimizzazione congiunta.

Nella figura 6.22 è riportato lo schema di figura 6.20 in cui sono stati esplicitati i fattori di scala dei due contributi relativi al codebook adattativo e stocastico. Le uscite dei filtri senza memoria sono date da

$$\hat{X}_{WLT}(n) = g_{LT} \cdot Z(n)$$
$$\hat{X}_{Wcbk}(n) = g_{cbk} \cdot W(n) \qquad (6.28)$$

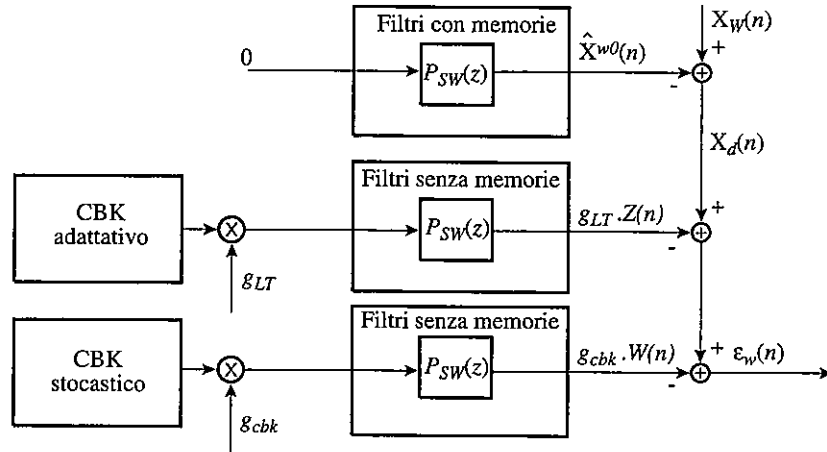



Fig. 6.22 - Schema del codificatore CELP per il calcolo con ortogonalizzazione.

in cui i segnali z(n) e w(n) rappresentano rispettivamente i contributi, filtrati, dei codebook adattativo e stocastico.

Come si è visto, la procedura di minimizzazione dell'energia dell'errore pesato consiste nella minimizzazione di

$$E = \sum_{n=1}^{L_s} \left(X_d(n) - g_{LT} \cdot Z(n) - g_{cbk} \cdot W(n) \right)^2$$

$$E = \sum_{n} \left(X_d^2(n) - g_{LT}^2 \cdot Z^2(n) - g_{cbk}^2 \cdot W^2(n) \right) +$$

$$- 2g_{LT} \sum_n X_d(n)\, Z(n) - 2g_{cbk} \sum_n X_d(n)\, W(n) + g_{cbk} g_{LT} \sum_n W(n)\, Z(n) \qquad (6.29)$$

in cui si nota che se $\sum_n W(n)\, Z(n) = 0$, e cioè se W(n) e Z(n) sono ortogonali, le derivate parziali rispetto a g_{ckb} e g_{LT} sono indipendenti.

La procedura di ortogonalizzazione consiste nel modificare il vettore W(n) affinché sia ortogonale a Z(n). Questo può essere realizzato ponendo:

$$W_{ORT}(n) = W(n) - \frac{\Psi}{\Gamma} \cdot Z(n) \qquad (6.30)$$

in cui

$$\Psi = \sum_{n=1}^{L_s} W(n)\,Z(n) \text{ e } \Gamma = \sum_{n=1}^{L_s} (Z(n))^2 \tag{6.31}$$

Con questa posizione si verifica facilmente che i coefficienti ottimi possono essere calcolati con le relazioni

$$g_{LT} = \frac{\sum_{n=1}^{L_s} X_d(n)\,Z(n)}{\sum_{n=1}^{L_s} (Z(n))^2}$$

$$g_{cbk} = \frac{\sum_{n=1}^{L_s} X_d(n)\,W_{ORT}(n)}{\sum_{n=1}^{L_s} (W_{ORT}(n))^2} \tag{6.32}$$

e l'energia del segnale errore, con i fattori di scala ottimi, vale

$$E_{ott} = \sum_{n=1}^{L_s} X_d^2(n) - \frac{\left(\sum_{n=1}^{L_s} X_d(n)\,Z(n)\right)^2}{\sum_{n=1}^{L_s} (Z(n))^2} - \frac{\left(\sum_{n=1}^{L_s} X_d(n)\,W_{ORT}(n)\right)^2}{\sum_{n=1}^{L_s} (W_{ORT}(n))^2} \tag{6.33}$$

Post-filtering

La tecnica del post-filtering è stata introdotta originariamente da Jyant come miglioramento della qualità nella codifica ADPCM [Ram85]. Successivamente è stata impiegata in numerosi schemi di codifica CELP ed è stata inserita nella raccomandazione G.728 relativa al codec LD-CELP così pure come nello standard a 8 kbit/s G.729.

Il principio è quello di operare una distorsione del segnale ricostruito tale che comporti una sagomatura del rumore di quantizzazione che lo renda meno

percepibile. In qualche misura il concetto è simile a quello alla base dell'introduzione della funzione di mascheramento spettrale, ma in questo caso l'operazione non avviene in fase di codifica bensì di decodifica, non ha quindi nessun effetto sulla determinazione dei parametri. Lo schema di impiego è riportato in figura 6.23 in cui $\hat{X}(n)$ costituisce l'uscita ricostruita dal decodificatore CELP.

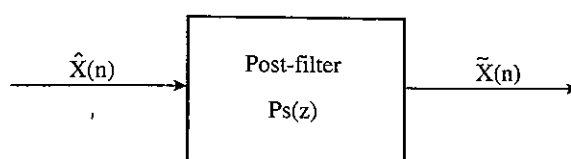



Fig. 6.23 - Schema di impiego del post-filter.

L'idea può essere discussa facendo riferimento alla figura 6.24a in cui lo spettro del segnale è costituito da due bande di ampiezza 30 dB e 10 dB rispettivamente, mentre il rumore di quantizzazione è a banda intera con valore di 15 dB. In queste condizioni il segnale nella prima banda è 15 dB sopra il rumore mentre nella seconda è 5 dB sotto.

Ipotizzando ora che la funzione di trasferimento del post-filter abbia la stessa caratteristica dello spettro del segnale vocale, lo spettro in uscita sarà quello riportato in figura 6.24b in cui l'SNR nelle due bande è rimasto invariato, mentre il segnale nella prima banda è ora 45 dB al disopra del rumore e quello della seconda è 5 dB. L'effetto complessivo è quindi quello di una minore percezione del rumore di quantizzazione. L'effetto secondario negativo, tuttavia, è quello di aver anche distorto il rapporto energetico tra le due bande di segnale.

Come funzione di post-filter si utilizza in pratica una funzione molto simile a quella inversa impiegata nel filtro di pesatura spettrale W(z) e cioè del tipo

$$A_{PS}(z) = \frac{A(z/\alpha)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i \alpha^i z^i}{1 - \sum_{i=1}^{p} a_i \beta^i z^i} \tag{6.34}$$

in cui i parametri α e β assumono valori compresi tra 0 e 1. Tipicamente $\alpha = 0.5$ e $\beta = 0.8$. La funzione di trasferimento con tali valori è riportata in figura 6.25.

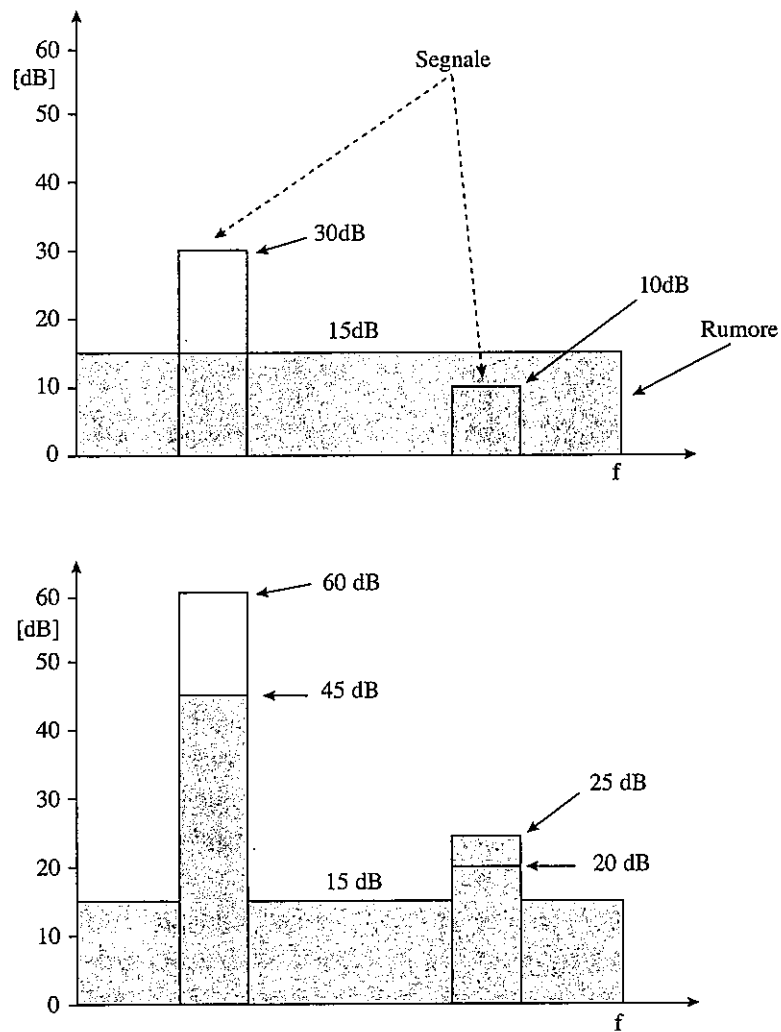



Fig. 6.24 - Schematizzazione del processo operato dal post-filter: (a) spettri del segnale e del rumore di quantizzazione all'ingresso del post-filter e (b) all'uscita del post-filter.

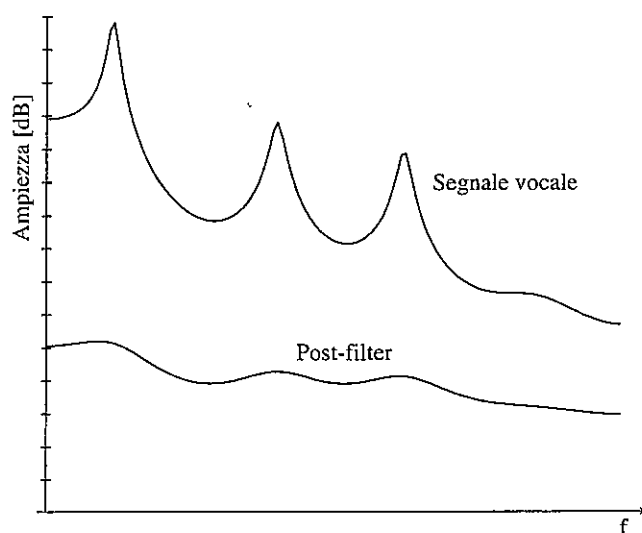



Fig. 6.25 - Inviluppo spettrale del segnale vocale ed andamento spettrale della funzione di trasferimento del post-filter con $\alpha = 0.5$ e $\beta = 0.8$.

6.5.4 Standard ITU-T G.729

Lo standard ITU-T G.729 è un codec basato sulla tecnica di codifica CELP. È stato proposto con il nome di CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Predictive) in considerazione della struttura del codebook impiegata. L'algoritmo di codifica opera a 8 kbit/s con un frame di 10 ms. Il ritardo di codifica è leggermente superiore (15 ms) in quanto l'analisi LPC è realizzata con un look-ahead di 5 ms e cioè per la stima dei parametri spettrali si considerano frame overlappati.

Lo schema del codificatore è riportato in figura 6.26. Il segnale di ingresso è preprocessato con un filtro passa alto con frequenza di taglio di 140 Hz al fine di evitare la presenza di disturbi in bassa frequenza. Sul segnale così filtrato sono calcolati i parametri del filtro di predizione a breve termine o filtro LPC. L'analisi, aggiornata ogni 10 ms (frame), è effettuata con una finestra lunga 30 ms spostata in avanti di 5 ms (look-ahead). Le finestre di analisi risultano quindi sovrapposte nel tempo (overlapped), secondo lo schema illustrato in figura 6.27.

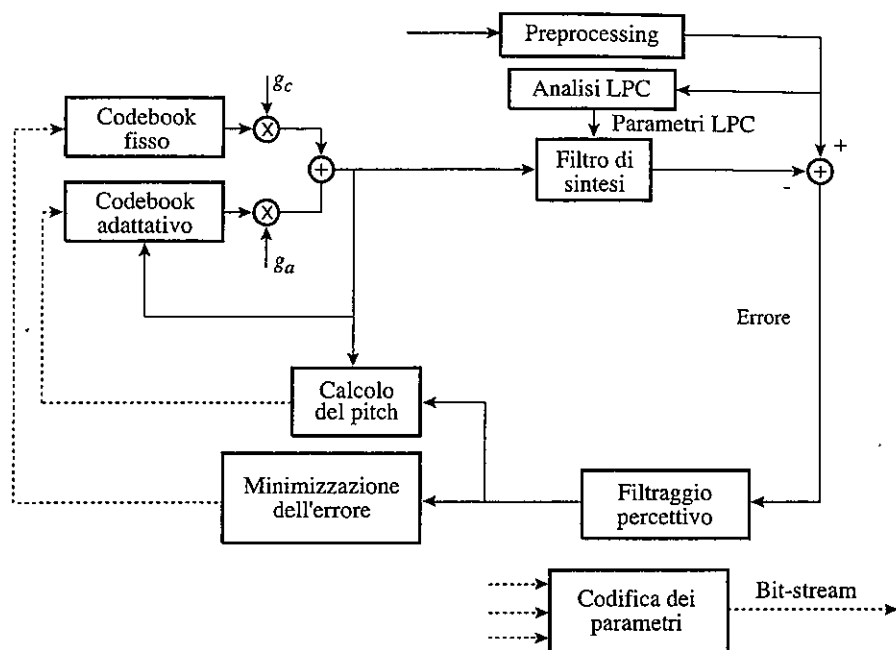



Fig. 6.26 - Schema del codificatore CS-ACELP.

La finestra di analisi impiegata smussa i valori dei campioni ai bordi, utilizzando una combinazione di finestra di Hamming e di una funzione coseno. L'analisi LPC, di ordine 10, è effettuata con il metodo dell'auto-correlazione introducendo una espansione delle formanti di 60 Hz ed aggiungendo rumore gaussiano di sottofondo a - 40dB. Entrambe le correzioni servono ad evitare la presenza di risonanze troppo pronunciate. Successivamente i coefficienti del filtro diretto a_i sono calcolati con l'algoritmo di Levinson-Durbin e sono trasformati nei coefficienti LSP utilizzando il metodo dei polinomi di Chebyshev prima descritto.

La quantizzazione dei coefficienti LSP è vettoriale predittiva tra set di coefficienti contigui. In pratica lo schema utilizzato per la quantizzazione è simile ad uno schema ADPCM in cui si utilizzano predittori di ordine 4.

I coefficienti a_i ottenuti per trasformazione inversa dai coefficienti $\hat{\omega}_i$, sono utilizzati direttamente per il secondo subframe, mentre i coefficienti impiegati nel primo subframe sono ottenuti per interpolazione lineare dei coefficienti LSP che, come si è visto precedentemente, costituiscono il dominio più efficace per effettuare tale operazione.

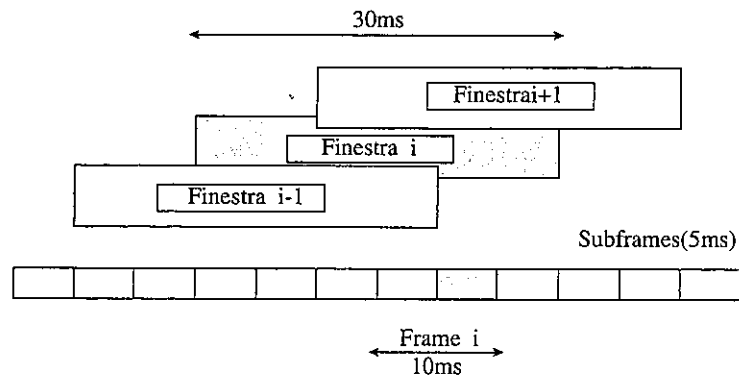



Fig. 6.27 - Schema di sovrapposizione delle finestre di analisi LPC.

Relativamente al filtro di pesatura percettiva, bisogna solo osservare che la funzione di trasferimento è del tipo

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \qquad (6.35)$$

in cui i due coefficienti γ_1 e γ_2 sono resi adattativi, frame per frame, sulla base della pendenza dello spettro del segnale vocale. Inoltre i coefficienti utilizzati sono quelli non quantizzati.

È interessante osservare che il segnale di ingresso pesato percettivamente, oltre a costituire il target per la ricerca nel codebook delle eccitazioni, è anche utilizzato per calcolare una stima del valore del pitch in catena aperta. Tale stima è calcolata massimizzando il coefficiente di autocorrelazione nell'intervallo tra 20 e 143 campioni. Al fine di evitare la scelta di multipli del pitch, la ricerca è effettuata in tre campi separati (tra 20 e 39, 40 e 79, 80 e 143) ed il candidato nell'intervallo più basso è favorito con una opportuna pesatura. Successivamente, il valore effettivo del pitch per ognuno dei due subframe da 5 ms è calcolato con procedura in closed loop nell'intorno della stima calcolata precedentemente. Per i motivi illustrati precedentemente, la ricerca del pitch nella seconda fase è realizzata nel dominio del segnale sovracampionato (con un fattore pari a 3).

Il codebook fisso, o stocastico, è del tipo "sparse" e cioè contiene solo quattro campioni non nulli sul subframe da 40 campioni. Gli impulsi possono

assumere i valori ± 1. I fattori di scala del codebook fisso e del codebook adattativo sono combinati in un vettore a due elementi e quantizzati vettorialmente. Più in particolare, il fattore di scala del codebook fisso è ottenuto applicando anche una tecnica predittiva a coefficienti costanti, al fine di minimizzare la dinamica del coefficiente stesso. In pratica quindi si calcola un errore di predizione tra il fattore di scala effettivo g_c e la versione predetta ottenuta considerando le versioni quantizzate dei fattori di scala dei quattro subframe precedenti. La predizione è fatta con un filtro a coefficienti costanti. L'errore di predizione è espresso in termini logaritmici e quindi può essere considerato come un fattore correttivo del fattore di scala predetto:

$$\lambda = g_c/g_p \qquad (6.36)$$

Tale fattore γ è combinato in un vettore a due elementi con il fattore di scala del codebook adattativo g_a e quantizzato vettorialmente usando un codebook da 128 parole (7 bit).

A titolo riassuntivo l'allocazione dei bit ai vari parametri trasmessi dal codec G.729 è riportata in tabella 6.4.

Parametro	Subframe 1	Subframe 2	Totale per frame
LSP			18
Ritardo cbk adattativo	8	5	13
CRC sul ritardo	1		1
Indice del cbk fisso	13	13	26
Segni del cbk fisso	4	4	8
Fattori di scala	7	7	14
Totale			80

Tab. 6.4 - Allocazione dei bit per il codec G.729.

Il decoder, come in ogni schema CELP, ha una complessità decisamente inferiore all'encoder ed effettua le operazioni di decodifica dei vari parametri trasmessi, nonché i filtraggi di sintesi sia a lungo che a breve termine. Infine, il segnale sintetizzato viene post-processato con un algoritmo di post-filtering che opera al fine di sagomare la distorsione introdotta. L'algoritmo è

particolarmente sofisticato in quanto considera non solo lo spettro a breve termine, ma anche quello a lungo termine, oltre ad un passo di controllo automatico del guadagno.

6.6 CODIFICA A VELOCITÀ VARIABILE

Nei capitoli precedenti si è parlato diffusamente di diverse tecniche di codifica sempre ipotizzando un segnale di ingresso ed uscita a velocità di trasmissione costante. Per l'ingresso questo significa campionare il segnale vocale ad una certa frequenza, tale da rappresentare la banda di segnale desiderata, e rappresentare ogni campione con un numero di bit-costante (ad esempio 16). Per l'uscita significa rappresentare ogni porzione temporale in questione (un campione oppure un frame) con un numero di bit costanti.

Questa configurazione è infatti di gran lunga la più impiegata in pratica e deriva principalmente dalla constatazione che il canale di trasmissione è spesso un canale con velocità di trasmissione costante.

Tuttavia, a seguito dello studio sempre più approfondito di codificatori vocali parametrici, è emerso in modo sempre più convincente che l'informazione associata al segnale vocale necessita di un flusso trasmissivo che non è costante nel tempo. In altre parole si può affermare che l'informazione necessaria a rappresentare accuratamente il segnale vocale è tempo variante o ancora che l'entropia a breve termine del segnale vocale varia considerevolmente nel tempo. Questa considerazione si spiega in modo semplice considerando una conversazione telefonica in cui un flusso trasmissivo è rappresentativo sia di tratti di voce attiva (talkspurt) sia di tratti di silenzio, quando uno dei due parlatori non parla, ma ascolta. È evidente che l'informazione necessaria a rappresentare con buona qualità il silenzio, o il rumore ambientale, è decisamente minore di quella necessaria a rappresentare il segnale vocale. Un esempio applicativo di questa constatazione si ha nello sviluppo del sistema DTX (Discontinuous Transmission) associato al sistema di comunicazione mobile GSM. In questa applicazione, il trasmettitore radio viene spento durante le pause della comunicazione con conseguente risparmio in termini di interferenza tra canali e di consumo delle batterie dei terminali mobili portatili. In questo caso il flusso informativo associato alle pause è addirittura azzerato.

L'esempio citato si riferisce ad un caso estremo di trasmissione on/off, tuttavia il concetto è estendibile anche allo stesso segnale attivo o talkspurt. Anche

per questi tratti, infatti, esistono porzioni del segnale per le quali le caratteristiche, ad esempio di correlazione, sono tali da consentire una modellizzazione efficiente e quindi un risparmio significativo in termini di bit trasmessi. È noto, ad esempio, [Kub93] che i tratti di segnale non vocalizzato possono essere riprodotti con buona fedeltà con una quantità di informazione minima, essenzialmente costituita dal livello e dalle caratteristiche spettrali. Viceversa, attacchi o transizioni tra suoni diversi, richiedono una grande quantità di informazione al fine di mantenere una buona qualità. Ne consegue che la qualità fornita da un codificatore a velocità di trasmissione costante dipende in larga misura dalla sua capacità di rappresentare correttamente quei segmenti che sono "più difficili" da codificare [Pas84].

Sebbene quindi la codifica a velocità variabile appaia una soluzione praticabile per ridurre la velocità di trasmissione media, in considerazione del meccanismo fisico di produzione del segnale vocale, molto spesso il canale fisico di trasmissione è un canale a velocità costante. In questo caso impiegare una codifica a velocità variabile determina un costo aggiuntivo che in qualche misura controbilancia il guadagno intrinseco di codifica. Il costo aggiuntivo è da intendersi nella necessità di impiegare degli opportuni registri tampone i quali, tra l'altro, determinano un aumento del ritardo di comunicazione. Mentre la maggiore complessità può essere bilanciata dalle migliori prestazioni, il ritardo aggiuntivo spesso non è tollerabile. Queste considerazioni sono alla base dello scarso successo ottenuto dalle tecniche di codifica a VR.

Tuttavia esistono applicazioni per le quali non esiste il vincolo di un canale di trasmissione a velocità costante. Una di queste, ad esempio, è store&forward nella quale un messaggio viene registrato e riascoltato in tempi diversi (segreterie telefoniche). Un'altra applicazione di recente interesse è quella delle comunicazioni mobili che utilizzano tecnica di accesso CDMA (Code Division Multiple Access) per la quale la trasmissione di flussi informativi a velocità variabile è particolarmente congegnale. In questi casi infatti la diversa velocità è trattata con diversi "spreading-factor" utilizzati per modulare il segnale numerico. Il vantaggio di operare a velocità variabile con tale tecnica di accesso è immediato, in quanto la riduzione di bit-rate si traduce direttamente in una riduzione dell'interferenza e pertanto in un aumento della capacità. Altre tecniche di accesso come Extended-TDMA [Kay92] e Packet Reservation Multiple Access (PRMA) [Goo90] possono analogamente beneficiare dell'impiego di codifiche a velocità variabile.

6.6.1 Classificazione del segnale vocale

Un elemento fondamentale per un codificatore a velocità variabile è costituito dal blocco di classificazione del segnale vocale. Più in generale, lo schema completo di un codificatore a velocità variabile può essere rappresentato dallo schema a blocchi di figura 6.28.

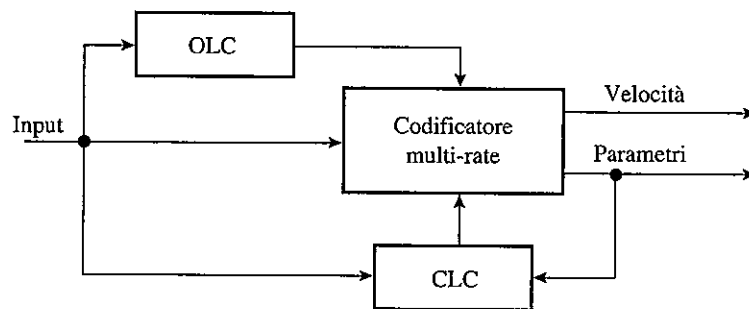



Fig. 6.28 - Schema a blocchi di un codificatore di sorgente a velocità variabile.

La struttura comprende un codificatore in grado di funzionare a velocità di trasmissione diverse e due blocchi di classificazione: uno inserito in catena aperta o Open Loop Classifier (OLC) ed uno in catena chiusa o CLC (Closed Loop Classifier). Lo scopo dei due blocchi è diverso. Mentre l'algoritmo inserito in catena aperta analizza il segnale vocale e lo classifica in base al suo contenuto intrinseco, il secondo algoritmo determina localmente la quantità di informazione necessaria al fine di garantire, con la trasmissione, un determinato grado di qualità. A tal fine, il secondo algoritmo utilizza misure oggettive tra il segnale da trasmettere e la sua versione sintetizzata localmente e pertanto considera sia il segnale di ingresso che quello ricostruito. La caratteristica principale di questo tipo di architettura è quella di tendere a fornire un segnale ricostruito di qualità costante con una velocità di trasmissione che viceversa è funzione del contenuto informativo e pertanto variabile.

In linea di principio si possono considerare diverse classificazioni del segnale vocale tese a massimizzare il guadagno di codifica. Una caratteristica comune è comunque quella che la classificazione non può prescindere dalla tecnica di codifica impiegata per ognuna delle classi identificate. In particolare se si fa riferimento al modello canonico di produzione del segnale vocale in cui si individuano tratti

vocalizzati e non, una semplice classificazione è quella riportata in figura 6.29 [Cel94].

segnale { attivo { vocalizzato / non vocalizzato ; inattivo }

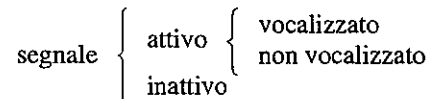



Fig. 6.29 - Classificazione ad anello aperto.

Secondo tale classificazione il segnale vocale è suddiviso in tratti attivi (o talkspurt) e tratti inattivi (quando il parlatore non parla). L'algoritmo impiegato per questa classificazione prende il nome di VAD (Voice Activity Detector) e sarà descritto nel capitolo successivo. I tratti attivi sono poi ulteriormente suddivisi in tratti vocalizzati e tratti non-vocalizzati. Questa classificazione fa riferimento al modello di produzione del segnale vocale descritto precedentemente.

Altre classificazioni più sofisticate sono state proposte in letteratura, come quella proposta da Gersho in [Pas84] in cui il segnale è suddiviso in quattro classi principali: noise, voiced, unvoiced e onset. Onset è considerato il primo frame voiced che segue un frame unvoiced. Inoltre i frame voiced sono ulteriormente sottoclassificati in full-band e low-pass in considerazione del contenuto energetico alle diverse frequenze. Questa classificazione è stata integrata con successo in uno schema di codifica basato sulla tecnica CELP.

Una classificazione ancora più diversificata è quella proposta in [Bat95] in cui il segnale vocale attivo è suddiviso nelle seguenti 5 classi: Noise, Onset, Steady-state, Decay, Periodic e Aperiodic.

Infine è interessante osservare come l'approccio di codifica basato su una classificazione possa essere esteso al caso limite in cui un sistema di comunicazione vocale è costituito da un classificatore in grado di riconoscere porzioni elementari del segnale vocale, come i fonemi o i difoni, (riconoscitore fonetico) e da un sintetizzatore che operi su una base di difoni opportuna (vedi fig. 6.30 [Car94]). In questo caso l'informazione da trasmettere è costituita dalla prosodia e dall'informazione dei difoni impiegati e si può stimare attorno ai 300 bit/s.

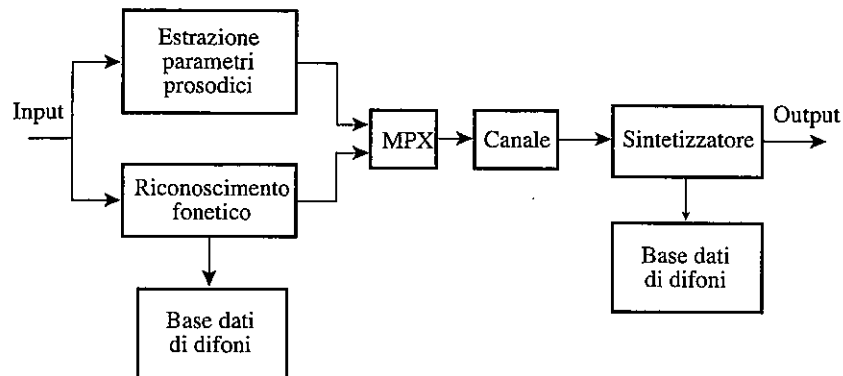



Fig. 6.30 - Schema di codifica realizzato con un riconoscitore ed un sintetizzatore.

6.6.2 VAD (Voice Activity Detection)

Un elemento comune ai vari classificatori presentati precedentemente è costituito dall'algoritmo di discernimento tra voce attiva e inattiva. Tale classificazione, che in linea di principio potrebbe sembrare relativamente semplice, risulta invece complessa quando si considerano condizioni di funzionamento più critiche. In particolare, la classificazione risulta delicata qualora l'indicazione debba essere fornita per lunghezze di frame molto corte (10 o 5 ms). In questi casi infatti, le pause di silenzio tipiche di alcuni fonemi (quali le plosive) possono essere interpretate come porzioni di silenzio. Un altro elemento che rende l'operazione difficile è la presenza di rumore ambientale che oltre ad alterare il livello del segnale vocale, determina anche una variazione delle caratteristiche spettrali del segnale in questione.

Un algoritmo che ha avuto particolare successo è quello impiegato nel sistema DTX del GSM che è stato standardizzato sia per il canale Full-rate [ETSI6.32] che per il canale Half-rate.

Lo schema a blocchi completo dell'algoritmo è riportato in figura 6.31. Le caratteristiche principali di questo schema sono quelle di un sistema progettato "fail-safe" e cioè polarizzato nei casi dubbi a fornire una indicazione di voce attiva piuttosto che silenzio. Inoltre, dovendo essere abbinato allo schema full-rate GSM, l'algoritmo utilizza i parametri calcolati nello schema RPE-LTP, al fine di minimizzare l'aumento di complessità dovuto al sistema DTX.

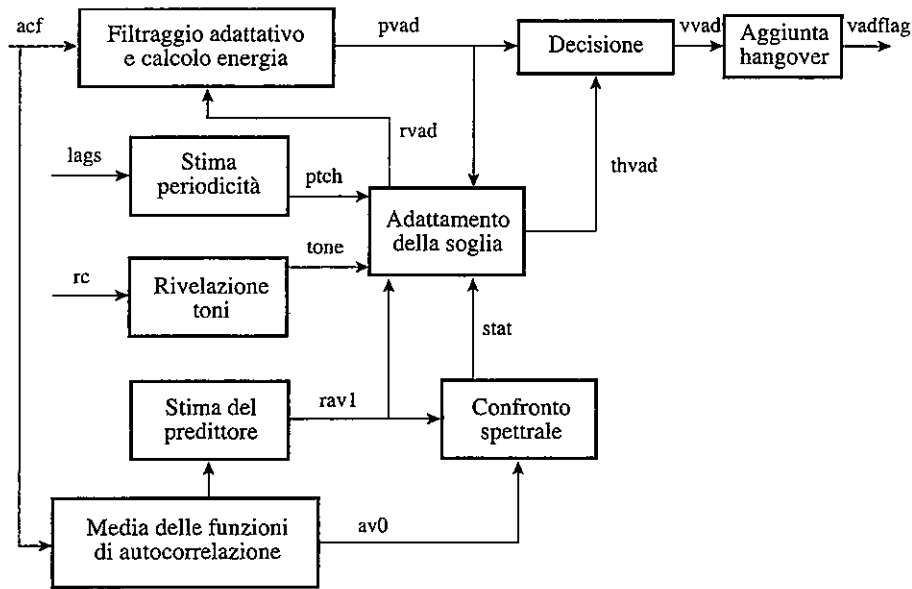



Fig. 6.31 - Schema a blocchi dell'algoritmo di VAD.

La decisione tra voce attiva e rumore di sottofondo è determinata confrontando l'energia del segnale di ingresso, opportunamente filtrato, con il valore di una soglia adattativa. Il VAD decide per voce quando l'energia è superiore a questa soglia.

Il segnale di ingresso è filtrato con un filtro di analisi di ordine 8 (lo stesso ordine impiegato nel codec RPE-LTP) i cui coefficienti sono calcolati a partire dai coefficienti di autocorrelazione del segnale di ingresso mediati su quattro frame consecutivi. Tale operazione di media consente di effettuare un filtraggio teso a ridurre il contenuto di rumore sovrapposto alla voce. Il risultato è quello di una discriminazione voce/rumore più affidabile. Al fine di ridurre al minimo l'aumento di complessità, l'operazione di filtraggio e determinazione dell'energia pvad sono effettuate con la relazione:

$$\text{pvad} = \text{rvad}(0) * \text{acf}(0) + 2 * \sum_{i=1}^{8} [\, \text{rvad}(i) * \text{acf}(i) \,] \tag{6.37}$$

in cui rvad(i) sono i coefficienti autocorrelazione dei coefficienti a(i) calcolati dall'algoritmo di VAD, mentre i valori acf(i) sono le funzioni di autocorrelazione del segnale di ingresso che sono già calcolate dall'algoritmo RPE-LTP.

L'operazione di adattamento della soglia di confronto è piuttosto elaborata ed è rappresentata dal diagramma di flusso riportato in figura 6.32.

Senza entrare in dettagli, tale operazione consiste nel modificare il valore della soglia per i tratti che si suppone siano relativi a voce non attiva. Tale evento è stimato dalle seguenti condizioni:

- l'energia del segnale è molto bassa (in tal caso la soglia è posta pari al valore minimo, plev);

oppure:

- lo spettro del segnale vocale è stazionario, il segnale non contiene una componente periodica e il segnale non contiene sinusoidi relative a toni di informazione della rete.

Nel primo caso i coefficienti del filtro adattativo non vengono aggiornati mentre nel secondo si. La stazionarietà del segnale è stimata calcolando il rapporto LHR (Likelihood Ratio) [Gra76] tra i coefficienti del filtro corrente e quello mediato sulle ultime quattro trame. Quando la distorsione spettrale LHR è inferiore ad una soglia fissa, il segnale è considerato stazionario. La determinazione della presenza di una componente periodica è agevolato dalla presenza dei valori del pitch calcolati nell'algoritmo RPE- LTP ogni 5 ms. La stima è effettuata considerando le relazioni tra i quattro valori relativi ad un frame da 20 ms.

La stima della presenza di toni di informazione è effettuata valutando il guadagno di predizione (rapporto tra l'energia del segnale e l'energia del segnale residuo). Quando il guadagno di predizione è inferiore ad una certa soglia (13.5 dB), si suppone non siano presenti toni. Tuttavia, in considerazione del fatto che il rumore veicolare può anche avere picchi di risonanza tali da consentire un forte guadagno di predizione, qualora il guadagno sia maggiore, si verifica anche che la frequenza del primo polo sia maggiore di 385 Hz, non esistendo toni di segnalazione a frequenza inferiore.

L'adattamento della soglia consiste nell'incremento o decremento di una quota percentuale secondo le relazioni riportate nel diagramma di flusso di figura 6.32.

Infine, onde evitare che le pause intersillabiche siano esclude dai periodi di attività, oppure che code di segnale vengano tagliate, la decisione di voce attiva viene mantenuta per un periodo pari a 5 trame (100 ms). Tale periodo prende il nome di hangover e viene aggiunto solo nei casi in cui il VAD abbia rivelato voce attiva per almeno 3 trame consecutive. Questo accorgimento

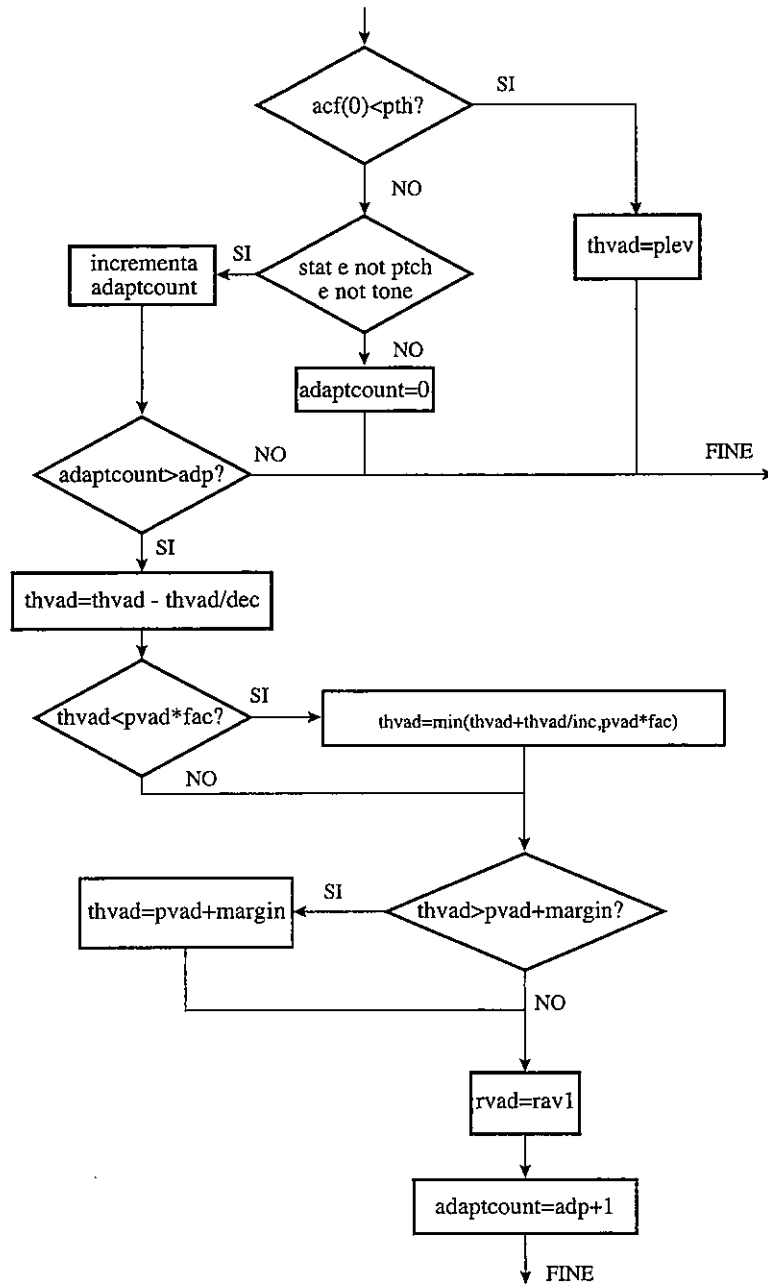



Fig. 6.32 - Diagramma di flusso dell'operazione di adattamento della soglia.

consente di evitare che disturbi occasionali presenti in tratti di silenzio prolunghino il periodo di attività rivelato erroneamente.

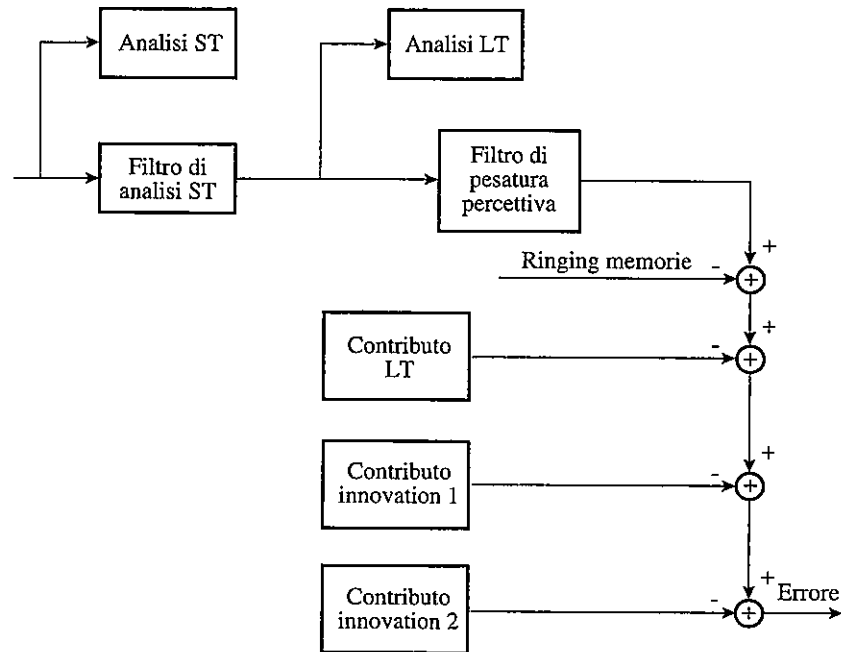



Fig. 6.33 - Schema a blocchi di un codificatore CELP a velocità variabile.

6.6.3 CELP a velocità variabile

Una particolare implementazione di codifica a velocità variabile è quella che impiega il codificatore CELP come algoritmo di codifica a diverse velocità di trasmissione. Se si fa riferimento alla classificazione semplice riportata in figura 6.29, si nota come l'algoritmo CELP si presti agevolmente ad essere impiegato. Infatti si può osservare come nel caso di segnale non vocalizzato, il contributo relativo alla correlazione a lungo termine possa essere eliminato con conseguente riduzione della velocità di trasmissione. Inoltre, strutturando opportunamente il contributo del codebook fisso come somma di più contributi in parallelo, si può concepire l'architettura di figura 6.33 [Cel94] in cui il secondo contributo di innovation è inserito solo in quei tratti in cui la qualità ottenuta con il singolo contributo non è sufficiente.

In questa particolare realizzazione, il codificatore CELP multi-rate è in grado di operare a 8 velocità differenti comprese tra un minimo di 400 bit/s ed un massimo di 16 kbit/s. Uno degli otto modi di funzionamento è inoltre impiegato per segnalare che nessun parametro è trasmesso, consentendo di annullare la velocità di trasmissione. Le velocità relative ad ogni modo di funzionamento ed i parametri trasmessi sono riassunti in tabella 6.5.

	MODO							
Parametri	1	2	3	4	5	6	7	8
Guadagno		+	+					
Parametri ST			+	+	+	+	+	+
Parametri LT						+	+	+
Innovation A				+	+		+	+
Innovation B					+			+
Velocità in kbit/s	0	0.4	3.2	8.5	12.5	7.2	12	16

Tab. 6.5 - Parametri trasmessi e velocità per i diversi modi di funzionamento.

Dalla tabella si nota che i primi tre modi non prevedono la trasmissione di parametri di eccitazione essendo infatti dedicati alla rappresentazione del rumore: in particolare il modo 2 per l'aggiornamento del livello, mentre il modo 3 anche per l'aggiornamento delle caratteristiche spettrali. I modi 4 e 5 sono invece impiegati per rappresentare i tratti non vocalizzati, in quanto i parametri del predittore a lungo termine non sono trasmessi. Infine gli ultimi tre modi sono impiegati per i tratti vocalizzati. La presenza di uno solo o due contributi di eccitazione è determinata dall'algoritmo CLC (vedi fig. 6.28) valutando le prestazioni ottenibili.

A titolo esemplificativo, la figura 6.34 riporta un tratto di forma d'onda relativo ad una conversazione e le relative velocità di trasmissione associate, ottenute con lo schema ora descritto. Risulta evidente da questo esempio che la velocità di trasmissione varia in modo considerevole da trama a trama.

La distribuzione statistica della velocità di trasmissione dipende ovviamente dal segnale vocale di ingresso ed in buona misura dal fattore di attività, oltre che dalle condizioni di rumore ambientale. La figura 6.35 riporta l'istogramma degli 8 modi di funzionamento per un insieme di quattro conversazioni e per una durata complessiva di diversi minuti. In particolare la

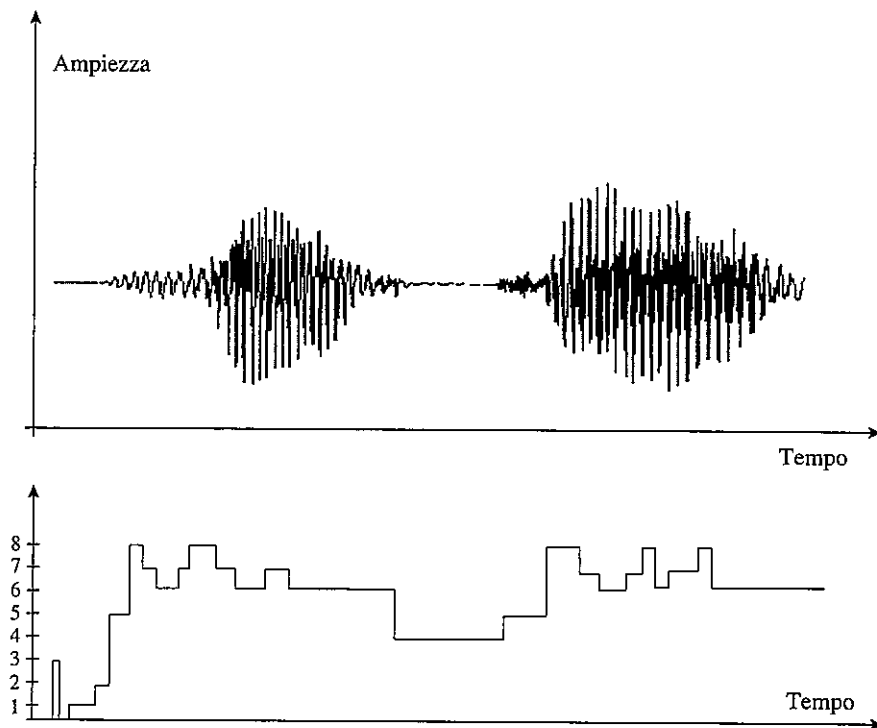



Fig. 6.34 - Andamento della velocità di trasmissione per un tratto di segnale vocale.

figura 6.35a si riferisce ad ambiente di ufficio e quindi con livello basso di rumore, mentre la figura 6.35b è relativa a conversazioni registrate in automobile e quindi con rumore ambientale più alto. È interessante osservare che nel passare da voce clean a voce con rumore, l'effetto è quello di aumentare la percentuale di modi di funzionamento a più alta velocità, sia per i tratti unvoiced che per i tratti voiced. Questo significa aggiungere il secondo contributo di innovation più frequentemente. Viceversa la percentuale di tempo classificato come voce inattiva rimane circa invariata. L'effetto è quello di aumentare la velocità di trasmissione media che passa da circa 6 kbit/s a 7.6 kbit/s. Parimenti, la qualità del segnale riprodotto è mantenuta. La tecnica realizza quindi quanto desiderato, nell'ottica della trasmissione a velocità variabile, e cioè aumenta la velocità per segnali complessi, mantenendo costante la qualità.

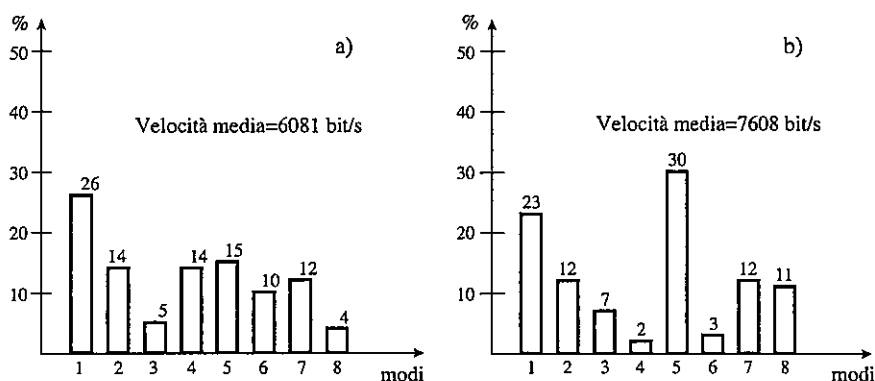



Fig. 6.35 - Percentuale di modi di trasmissione nella codifica a velocità variabile. a) ambiente di ufficio - b) ambiente automobile.

6.6.4 Controllo della velocità di trasmissione

Nel paragrafo precedente si è parlato del concetto di codifica a velocità variabile come diretta conseguenza delle caratteristiche intrinseche della sorgente. Considerando un sistema di comunicazione completo, tuttavia si possono identificare altri parametri di ingresso che possono determinare l'esigenza di modificare la velocità di trasmissione.

In generale si può immaginare un sistema di trasmissione che comprenda diverse sorgenti di informazioni, non necessariamente solo segnale vocale ma anche segnalazione o segnale video o altri [Ber93], sovrapposte sullo stesso canale di trasmissione (fig. 6.36).

Tale sistema comprende un blocco VRCU (Variable Rate Control Unit) che, in considerazione delle diverse informazioni raccolte, frame per frame, determina la velocità di trasmissione ottima per la specifica sorgente e la quantità di protezione necessaria.

I parametri di ingresso a tale blocco sono i seguenti:

- Sorgente: comprende l'informazione sulla velocità di trasmissione ottima per il frame in questione. È funzione del segnale di ingresso.
- Utente: determina la qualità di comunicazione in considerazione delle esigenze dell'utente (ad esempio banda stretta o banda larga). È un'informazione che tipicamente varia con meno frequenza, potrebbe essere aggiornata per ogni conversazione.

- Canale: determina la quantità di ridondanza necessaria a proteggere l'informazione in considerazione delle condizioni di canale. È un parametro che cambia frequentemente nel tempo soprattutto nel caso di canali radio.
- Sistema: comprende le informazioni sul traffico. Nel caso di comunicazioni cellulari tale informazione è variabile nel tempo e dipende dalla posizione del terminale mobile come pure dal traffico in una cella.

Un sistema di codifica che consenta la flessibilità di modificare la velocità di trasmissione non solo in considerazione delle caratteristiche del segnale di ingresso, ma anche in considerazione di parametri esterni, deve avere opportune caratteristiche. In particolare, se si vuole modificare la velocità lungo la catena trasmissiva, il sistema di codifica deve consentire la decodifica del segnale anche quando parte dell'informazione trasmessa non arriva a destinazione (codifica embended).

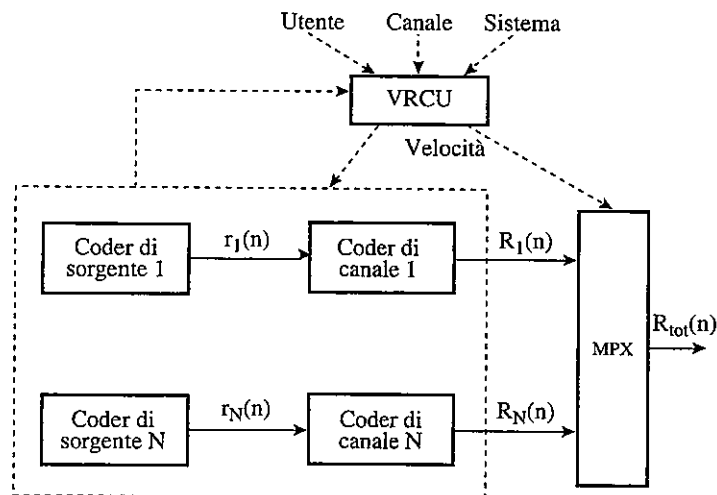



Fig. 6.36 - Sistema di trasmissione con controllo della velocità.

6.6.5 Le tecniche di codifica embedded

Nei casi in cui sia importante che la riduzione di velocità di trasmissione sia realizzata in un punto generico della catena trasmissiva, allora il sistema di codifica deve essere un sistema "embedded." Un codificatore embedded infatti

è in grado di produrre un flusso di bit che contiene annidato (embedded) un flusso a velocità inferiore tale da poter essere impiegato al ricevitore per effettuare la sintesi del segnale, seppur con qualità inferiore.

Un esempio di tecnica di codifica embedded è costituito dallo standard ITU-T G.727 [ITU-T G.727] che impiega la tecnica ADPCM con velocità 16,24,32 e 40 kbit/s. In questo caso il segnale errore di predizione è quantizzato con 5 bit/campione, ma solo 2 sono utilizzati localmente al trasmettitore per generare il segnale predetto ed il segnale ricostruito che è usato per calcolare i coefficienti del predittore. Tale accorgimento riduce le prestazioni ottenibili alla velocità massima (in quanto la stima dei parametri impiegati è fatta con un'informazione minore), ma parimenti consente di avere una degradazione della qualità graduale e non drammatica nel ridurre la velocità di trasmissione.

Lo stesso principio è stato applicato in [Dro91] ad uno schema di codifica di tipo CELP in cui i contributi aggiuntivi alla velocità di trasmissione sono costituiti da rami di eccitazione aggiuntivi in parallelo.

Analogamente in [Cel96] si propone uno schema embedded per codifica di segnali audio in cui la velocità di trasmissione è variabile a passi di 2 kbit/s da un minimo di 6 kbit/s ad un massimo di 64 kbit/s. La variazione di velocità di trasmissione è accompagnata da un aumento della banda di segnale audio trasmesso che raggiunge i 20 kHz a 64 kbit/s. L'elemento principale di questa architettura è costituito dalla BMU (Bit Manipulation Unit), la quale consente di operare sul flusso di bit trasmessi in modo tale da ridurre la velocità di trasmissione in considerazione delle esigenze esterne (fig. 6.37).

La struttura del codificatore è tale che il segnale audio in ingresso è codificato da un algoritmo core alla velocità base di 6 kbit/s mentre l'errore di quantizzazione (nel dominio del tempo o della frequenza a seconda dell'algoritmo core usato) viene codificato da stadi di enhancement i quali producono pacchetti di informazione da 2 kbit/s. La velocità complessiva è di 64 kbit/s ed i pacchetti sono organizzati come illustrato in figura 6.37b. La BMU è quindi in grado di ridurre la velocità di trasmissione a passi di 2 kbit/s, tralasciando i pacchetti secondo un ordine di importanza percettiva. In particolare questo comporta che al diminuire della velocità di trasmissione, anche la banda del segnale audio riprodotto diminuisce.

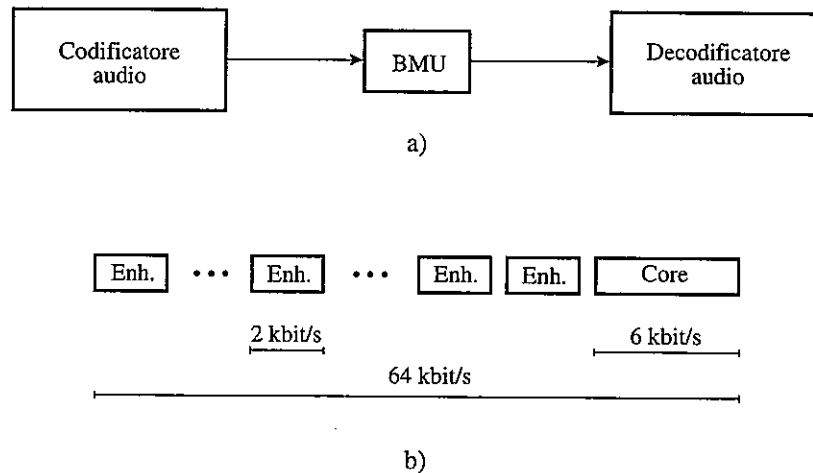



Fig. 6.37 - a) Schema di comunicazione con BMU e b) organizzazione dei pacchetti trasmessi.

6.7 LA TECNICA PWI

La tecnica di codifica PWI (Prototype Waveform Interpolation) si presta particolarmente ad essere impiegata per tratti di segnale vocale molto stazionario e cioè tipicamente per tratti di segnale fortemente vocalizzato. Non si presta invece ad essere usata per i tratti non vocalizzati dove non esiste un prototipo di forma d'onda di riferimento. La tecnica è stata impiegata quindi in codificatori multi modo che impiegano tecniche diverse per tratti di segnale diversi ed in particolare per codificatori a velocità variabile.

L'algoritmo PWI genera il segnale di eccitazione del filtro di sintesi LPC interpolando forme d'onda "prototipi" di lunghezza pari al periodo del pitch. Lo strumento base impiegato è costituito spesso dalla DFT pitch sincrona.

È stato più volte ricordato che il segnale residuo di predizione relativo ai tratti vocalizzati presenta una struttura quasi periodica con periodo pari al pitch. È quindi possibile suddividere un determinato frame di segnale vocalizzato in una sequenza di Nr segmenti di lunghezza P, con P periodo del pitch, come indicato in figura 6.38.

È evidente che in generale un frame non conterrà un numero intero di segmenti e pertanto si considerano anche campioni del frame precedente. Con questa operazione si producono Nr vettori di lunghezza P che possono essere

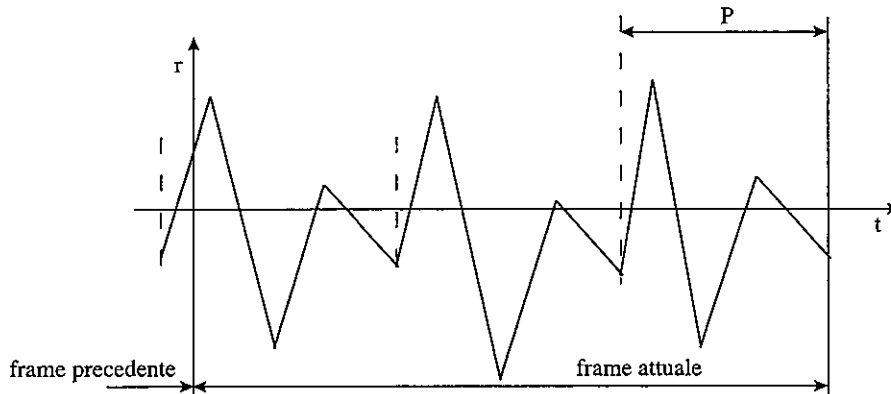



Fig.6.38 - Suddivisione di un frame in sottosegmenti di lunghezza pari al pitch.

ordinati in una matrice R nella quale ogni colonna contiene i campioni di segnale residuo relativo ad un periodo del pitch.

$$R = \begin{bmatrix} r(0,0) & r(0,1) & \dots & r(0,N_r-1) \\ r(1,0) & r(1,1) & \dots & r(1,N_r-1) \\ \dots & \dots & \dots & \dots \\ r(P-1,0) & r(P-1,1) & \dots & r(P-1,N_r-1) \end{bmatrix} \qquad (6.38)$$

Il numero di colonne dipende dal valore del pitch e dalla lunghezza del frame che, qualora il ritardo di trasmissione non sia un vincolo forte, può anche essere molto lunga, in quanto la tecnica è impiegata per tratti molto stazionari, tratti in cui anche i parametri spettrali cambiano poco.

Diverse soluzioni sono state proposte in letteratura su come ricostruire l'informazione a partire da questa matrice [Kle91, Sho93, Tan94, Fes95]. Kleijn [Kle91] propone di quantizzare e trasmettere un prototipo di forma d'onda e generare il segnale mancante con tecniche di interpolazione lineare. La soluzione proposta da Shoham [Sho93] con il nome di TFI (Time-Frequency Interpolation) prevede di trasformare con DFT la prima colonna, eventualmente effettuare una sagomatura spettrale, quantizzare e trasmettere i moduli. Il segnale al ricevitore è ricostruito utilizzando una antitrasformata DFT modificata, in cui la fase è ottenuta per interpolazione lineare assumendo lineare la traiettoria di variazione del pitch. La soluzione proposta da Tanaka [Tan94] considera una DFT bidimensionale ottenuta

applicando prima una DFT alle colonne della matrice R e poi una DFT alle righe ottenendo la matrice R2. In tale matrice, ogni riga ha il significato di andamento in frequenza di ogni componente di frequenza del periodo di pitch.

$$R_2 = \begin{bmatrix} r_2(0,0) & r_2(0,1) & \dots & r_2(0,N_r-1) \\ r_2(1,0) & r_2(1,1) & \dots & r_2(1,N_r 1) \\ \dots & \dots & \dots & \dots \\ r_2(P-1,0) & r_2(P-1,1) & \dots & r_2(P-1,N_r-1) \end{bmatrix} \tag{6.39}$$

Considerando che le fluttuazioni del pitch ad alta frequenza non possono essere conservate con il metodo di interpolazione tipico della PWI, tali fluttuazioni sono annullate ponendo a zero le colonne centrali della matrice R2. La matrice così ottenuta viene poi antitrasformata per righe e la prima colonna costituisce il prototipo che deve essere trasmesso. Tale prototipo è codificato e trasmesso ad una velocità di trasmissione di circa 2 kbit/s. In [Fes95] infine, il vettore prototipo nel dominio della frequenza è quantizzato sfruttando le proprietà delle bande critiche e la caratteristica di quasi linearità della fase.

7

CODIFICA NEL DOMINIO DELLA FREQUENZA

7.1 GENERALITÀ SULLA CODIFICA NEL DOMINIO DELLA FREQUENZA

Le codifiche di forma d'onda e per modelli precedentemente presentate lavorano essenzialmente nel dominio del tempo. In tal modo risulta difficile estrarre alcune ridondanze presenti nel segnale che richiedono un'analisi in frequenza dello stesso. In particolare, gli aspetti su cui è possibile lavorare con un'analisi in frequenza sono:

- diversa struttura del segnale vocale. Lo spettro del segnale vocale può essere suddiviso fondamentalmente in due bande. Nella banda al di sotto dei 2 kHz si trovano le componenti relative ai suoni vocalizzati che richiedono un'estrema cura nella codifica. Nella banda superiore dello spettro si trova il contributo dei suoni non vocalizzati che, essendo approssimabili a rumore, sopportano codifiche meno curate senza un proporzionale degradamento della qualità percepita;
- differente dinamica delle componenti spettrali del segnale. All'aumentare della frequenza, la dinamica del segnale tipicamente diminuisce. Questo permette di adottare quantizzatori che, fissa l'ampiezza del quanto, abbiano estremi di saturazione (e quindi numero di livelli) che variano in funzione della frequenza del segnale;
- effetti di mascheramento presenti nell'apparato uditivo. A causa dell'impossibilità di percepire dettagli del segnale che si trovano al disotto della soglia di udibilità, soglia che viene influenzata da

fenomeni di mascheramento, è possibile sia adottare quanti tali che il rumore di quantizzazione venga mascherato dal segnale, sia eliminare alcune componenti armoniche del segnale di dinamica modesta, che comunque non verrebbero percepite.

In ogni caso è necessario rappresentare il segnale in un dominio differente da quello del tempo. Considerando per semplicità una rappresentazione in frequenza, la trasformazione può avvenire in due modi distinti. Dato un segnale discreto x(n) e considerato un blocco di campioni pesato tramite una finestra h(n), la sua trasformata di Fourier è definita come

$$X_n(j\omega) = \sum_{m=-\infty}^{\infty} h(n-m)\, x(m)\, e^{-j\omega m} \tag{7.1}$$

Questa trasformata è funzione di due variabili: l'indice temporale n e la pulsazione ω. Se si tiene fissa la pulsazione $\omega = \omega_0$, la trasformata è interpretabile come l'uscita di un filtro con risposta impulsiva h(n) ed ingresso pari al segnale modulato $x(n)\, e^{-j\omega_0 n}$, cioè

$$X_n(j\omega_0) = h(n) \otimes \left[\, x(m)\, e^{-j\omega_0 m} \,\right] \tag{7.2}$$

La modulazione introdotta nella trasformata ha lo scopo di riportare nell'intorno dell'origine una porzione del segnale, la cui banda è fissata dal filtro adottato. Al variare di ω_0, si suddivide lo spettro del segnale in blocchi, analogamente a quanto avviene per mezzo di un banco di filtri passa banda (*codifica per sottobande*). Il flusso continuo emesso da ciascun filtro è poi codificabile con le tecniche tipiche dei codificatori di forma d'onda (vedi par. 7.2).

Viceversa, se si tiene fisso l'indice n = n0, Xn0(jω) è interpretabile come la trasformata del blocco di campioni [h(n0 m) x(m)] centrato sull'istante $t = n_0 \times T$

$$X_{n_0}(j\omega) = F\left[\, h(n_0 - m)\, x(m) \,\right] \tag{7.3}$$

Trasformando ciascun blocco del flusso numerico, si ottiene una rappresentazione in frequenza, i cui coefficienti costituiscono la codifica del segnale stesso (*codifica per trasformate*) (vedi par. 7.3).

Sebbene le due tecniche siano simili nell'impostazione (fig. 7.1), codifiche per trasformate eseguite su blocchi estesi di campioni permettono

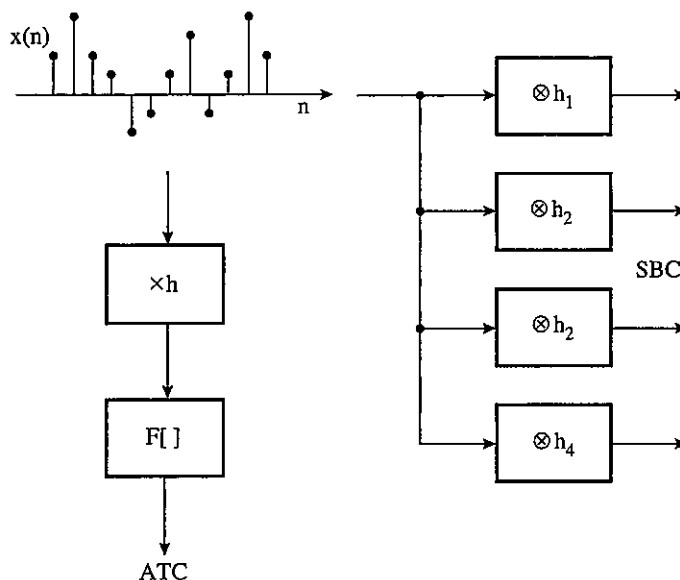



Fig. 7.1 - Differenza tra codifica per sottobande e per trasformate.

una migliore risoluzione in frequenza rispetto a codifiche per sottobande. D'altra parte, all'aumentare della dimensione dei blocchi utilizzati nelle trasformate peggiora la risoluzione temporale ottenibile. Per tali motivi, i maggiori rapporti di compressione si ottengono adottando tecniche ibride, anche se a prezzo di una maggiore complessità computazionale.

7.2 CODIFICA PER SOTTOBANDE

7.2.1 Codifica per sottobande di forma d'onda

Concettualmente, la codifica per sottobande (sub-band coding: SBC) è ottenibile filtrando innanzitutto il segnale tramite un banco di M filtri passa banda e riportando poi ciascuna banda nell'intorno dell'origine, tramite una sua modulazione (fig. 7.2). Un primo problema da affrontare è, quindi, quello di determinare l'ampiezza di ciascuna sottobanda, dato che riducendo l'ampiezza delle sottobande si migliorano le prestazioni del codificatore, ma se ne aumenta la complessità. La suddivisione in sottobande può avvenire secondo più criteri. Innanzitutto è possibile utilizzare bande equispaziate e non

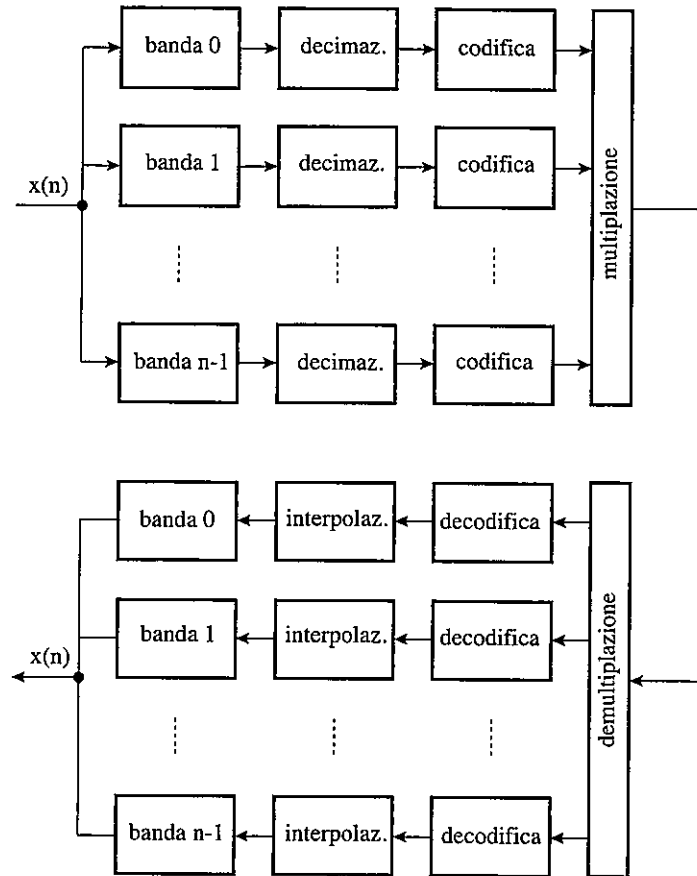



Fig. 7.2 - Codifica per sotto-bande.

equispaziate (fig. 7.3). Considerando la distribuzione spettrale di potenza del segnale, la prima soluzione viene preferita nel caso di distribuzione omogenea (es.: segnale audio musicale). In tal caso, infatti non vi è ragione di privilegiare porzioni dello spettro nei confronti di altre.

Nel caso di segnale vocale, invece, le bande, tipicamente, non sono equispaziate, ma si tenta di far coincidere le frequenze centrali di ciascuna sottobanda con quelle delle formanti. In tal modo vengono riprodotte al meglio le componenti del segnale che hanno un maggiore impatto sulla qualità di codifica. Le sottobande ottenute in questo caso risultano essere di ampiezza maggiore al crescere della frequenza. Nel caso di spettro particolarmente disomogeneo è possibile addirittura utilizzare bande che risultino non contigue.

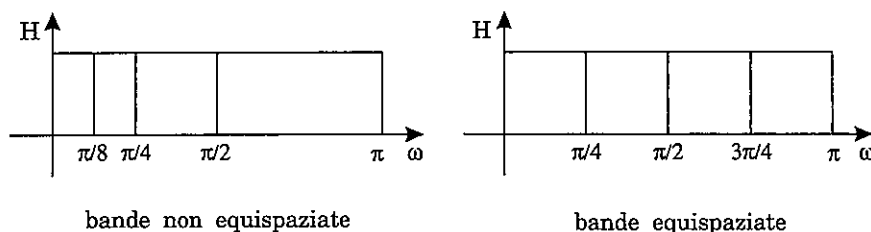



Fig. 7.3 - Possibili scelte per l'ampiezza delle sottobande.

Per il segnale vocale, ad esempio, tale soluzione sarebbe praticabile rinunciando a riprodurre le componenti spettrali poste al di fuori delle formanti, caratterizzate da ampiezze modeste e, quindi, trascurabili.

Anche nel caso si consideri la risoluzione in frequenza dell'apparato uditivo, le bande utilizzate non sono equispaziate (vedi par. 7.3.3), dato che le bande critiche hanno ampiezza crescente all'aumentare della frequenza.

Nel seguito del paragrafo verrà considerato solamente il caso di bande equispaziate. In tal caso, indicando con ω_x la banda complessiva del segnale, ciascuna sottobanda risulterà di ampiezza pari a $\omega_B = \omega_x/M$. Ipotizzando di lavorare sulla rappresentazione numerica del segnale, i filtri passa banda richiesti risulteranno digitali. Tralasciando per il momento il problema della traslazione della sottobanda nell'intorno dell'origine, si nota che sull'uscita di ciascun filtro è possibile eseguire una decimazione. Infatti, essendo l'ampiezza della banda pari a ω_B, la frequenza di campionamento richiesta è corrispondente ad una pulsazione $2\omega_B$, mentre il segnale è campionato per una pulsazione $\omega_x = M \times \omega_B$. Il fattore di sovracampionamento è, quindi pari a M. Di conseguenza, la sequenza decimata y(n), corrispondente ad un periodo di campionamento $T' = M\,T$, si ottiene dal segnale continuo $x_c(t)$ come

$$y(n) = x_c(nT') = x_c(nMT) = x(nM) \tag{7.4}$$

Considerando anche l'effetto dei filtri passa banda, le sequenze di uscita (non decimate e decimate) possono essere espresse come

$$w(n) = \sum_{k=-\infty}^{\infty} h(k)\, x(n-k)$$

$$y(m) = \sum_{k=-\infty}^{\infty} h(k)\ x(M \times m - k) = \sum_{n=-\infty}^{\infty} h(M \times m - n)\ x(n) \tag{7.5}$$

Tali relazioni sono esprimibili in frequenza come

$$Y\left(e^{j\omega T'}\right) = \frac{1}{M}\sum_{k=0}^{M-1} X\left(e^{j\frac{\omega T' - 2\pi k}{M}}\right) \tag{7.6}$$

In ricezione le sequenze vanno interpolate, innanzitutto inserendo campioni nulli al di fuori degli istanti di campionamento

$$w(m) = \begin{cases} x\left(\frac{m}{M}\right), & m = \pm M, \pm 2M, \dots \\ 0\,, & \text{altrimenti} \end{cases} \tag{7.7}$$

e poi interpolando la nuova sequenza tramite un filtraggio numerico passa basso con frequenza di taglio corrispondente all'ampiezza della sottobanda

$$y(m) = \sum_{k=-\infty}^{\infty} h(m-k)\ x\left(\frac{k}{M}\right) = \sum_{n=-\infty}^{\infty} h(m - M \times n)\ x(n) \tag{7.8}$$

Tale relazione in frequenza è pari a

$$Y\left(e^{j\omega T'}\right) = \frac{1}{M}\sum_{k=0}^{M-1} X\left(e^{j\frac{\omega T' - 2\pi k}{M}}\right) \tag{7.9}$$

Per i dettagli implementativi sulla struttura dei filtri necessari per la conversione della frequenza di campionamento si rimanda in [Appendice D].

Il ricampionamento, grazie alle repliche da esso introdotte, esegue automaticamente la necessaria traslazione del segnale nell'intorno dell'origine. Infatti, è sufficiente eliminare le repliche non necessarie tramite un filtro passa basso con frequenza di taglio ω_B. C'è, però, da tener presente che, a secondo che si consideri una sottobanda di ordine pari o dispari, la decimazione fornisce una sua replica esatta o rovesciata nell'intorno dell'origine (fig. 7.4). Nel

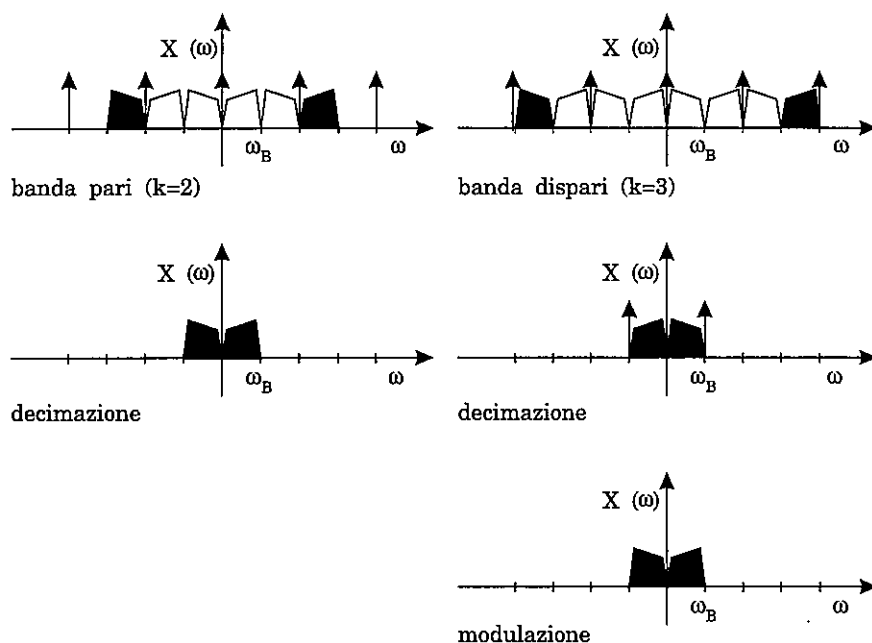



Fig. 7.4 - Decimazione di sottobande di ordine pari e dispari.

secondo caso è possibile ripristinare la corretta orientazione tramite una modulazione con un segnale sinusoidale a frequenza pari a metà della frequenza di campionamento. Numericamente ciò si ottiene moltiplicando i campioni per la sequenza (-1)n, cioè cambiando di segno ad un campione su due.

Ciascuna banda viene, infine, codificata tramite tecniche di forma d'onda. Nel caso di bande piuttosto piccole, dato che esse risultano essere campionate a frequenza molto prossima alla frequenza di Nyquist, non risulta utile adottare codifiche predittive. Viceversa, il vantaggio principale della codifica per sottobande risiede nella possibilità di poter quantizzare ciascuna banda con un numero di bit Rk variabile. Trascurando gli aspetti percettivi, trattati in un successivo paragrafo, per fissare le caratteristiche di quantizzazione nelle diverse sottobande si possono adottare più criteri, a volte tra loro combinabili, mentre altre volte tra loro contrastanti. Ad esempio:

- è possibile tentare di minimizzare l'errore globale di codifica per un fissato flusso numerico;

- utilizzando sempre quanti della stessa ampiezza, è possibile fissare gli estremi di saturazione di ciascuna sottobanda (e quindi il numero di bit) in maniera che si riducano al diminuire della potenza del segnale (tipicamente a frequenza crescente) (in maniera simile a quanto fatto nella codifica APCM);
- se si tenta di rendere uniforme il rapporto segnale rumore in frequenza, è possibile diminuire l'ampiezza dei quanti (e quindi il rumore) per le componenti ad energia inferiore (tipicamente a frequenza crescente) (in maniera simile a quanto fatto nella codifica NFC);
- sfruttando le caratteristiche della sorgente, nel caso di segnale vocale è possibile assegnare un numero di bit inferiore alle bande a frequenza maggiore (interessate alla produzione di suoni non vocalizzati e, quindi, più simili a rumore) ed un numero maggiore per le bande inferiori (legate alla ricostruzione del pitch e delle formanti) (es.: codifica ITU-T G.722).

Ipotizziamo di voler fissare la distribuzione di bit tra sottobande in modo che venga minimizzato il rumore di quantizzazione complessivo, per un fissato flusso numerico. Se si indica con Rk il numero di bit utilizzati nella sottobanda k-esima, con fs la frequenza di campionamento ed ipotizzando M sottobande equispaziate, il flusso totale generato nella codifica SBC è pari a

$$f_c = f_s \sum_{k=0}^{M-1} R_k \tag{7.10}$$

Per quanto riguarda la potenza dell'errore di quantizzazione, ipotizzando sottobande non sovrapposte risulta

$$\sigma_x^2 = \sum_{k=0}^{M-1} \sigma_k^2 \tag{7.11}$$

dove con σ_x^2 è indicata la potenza complessiva del segnale, mentre con σ_k^2 la potenza che cade nella sottobanda k-esima. Ricordando dalla 2.45 che la componente granulare dell'errore di quantizzazione vale

$$e_g^2 = \varepsilon_x^2 \, \sigma_x^2 \, 2^{-2R_k} \tag{7.12}$$

e nell'ipotesi $\varepsilon_k^2 = \varepsilon_x^2 = \varepsilon^2$, si ricava che la potenza complessiva dell'errore di quantizzazione è pari a

$$\varepsilon_g^2 = \varepsilon^2 \sum_{k=0}^{M-1} \sigma_k^2 \, 2^{-2R_k} \tag{7.13}$$

Il problema è minimizzare tale grandezza rispetto alle incognite Rk, con il vincolo che il flusso numerico

$$M R = \sum_{k=0}^{M-1} R_k \tag{7.14}$$

si mantenga costante. In questa espressione, R è il numero medio di bit per campione. Utilizzando i moltiplicatori di Lagrange, ciò equivale a minimizzare la funzione non vincolata

$$L = \varepsilon^2 \sum_{k=0}^{M-1} \sigma_k^2 \, 2^{-2R_k} - \lambda \left(M R - \sum_{k=0}^{M-1} R_k \right) \tag{7.15}$$

rispetto alle incognite Rk, con λ parametro arbitrario. Il minimo si ottiene imponendo che

$$\frac{\partial}{\partial R_k} \left[\varepsilon^2 \sum_{k=0}^{M-1} \sigma_k^2 \, 2^{-2R_k} - \lambda \left(M R - \sum_{k=0}^{M-1} R_k \right) \right] = 0 \tag{7.16}$$

da cui si ricavano gli Rk in funzione di λ

$$R_k = \frac{1}{2} \log_2 \left(2 \, \varepsilon^2 \log_e 2 \right) + \frac{1}{2} \log_2 \left(\frac{\sigma_k^2}{\lambda} \right) \tag{7.17}$$

La cercata allocazione ottima dei bit è, quindi, data da

$$R_{k,ott} = R + \frac{1}{2} \log_2 \left[\frac{\sigma_k^2}{\left(\prod_{j=0}^{M-1} \sigma_j^2 \right)^{1/M}} \right] \tag{7.18}$$

Sostituendo tale espressione in quella dell'errore di quantizzazione, si ottiene la potenza dell'errore di codifica

$$e^2_{min} = M\ \varepsilon_k^2\ 2^{-2R} \left[\prod_{j=0}^{M-1} \sigma_j^2\right]^{1/M} \tag{7.19}$$

Per valutare le prestazioni di questa codifica, è possibile definire il guadagno di codifica per sottobande G_{SBC}, come il miglioramento del rapporto segnale rumore ottenuto rispetto a quanto ottenuto con la codifica PCM. Sempre nell'ipotesi di $\varepsilon_k^2 = \varepsilon_x^2 = \varepsilon^2$, questo è pari a

$$G_{SBC} = \frac{\sigma_x^2\ 2^{-2R}}{\sum_{k=0}^{M-1} \sigma_k^2\ 2^{-2R_k}} \tag{7.20}$$

Si nota innanzitutto che tale guadagno può essere maggiore di uno solamente nel caso di segnale a spettro non piatto. Sostituendo poi in esso l'espressione degli Rk ottimi, il massimo guadagno di codifica che si ottiene è pari a

$$G_{SBC,max} = \frac{\frac{1}{M}\sum_{j=0}^{M-1} \sigma_j^2}{\left(\prod_{j=0}^{M-1} \sigma_j^2\right)^{1/M}} \tag{7.21}$$

Esempio [Jay84]: si consideri la densità spettrale di potenza riportata in figura 7.5, che nella metà inferiore e superiore della banda del segnale ha varianza pari a

$$\sigma_1^2 = \frac{16}{17}\ \sigma_x^2;\ \ \sigma_2^2 = \frac{1}{17}\ \sigma_x^2 \tag{7.22}$$

Fissato un valor medio di bit per campione R = 3, l'allocazione di bit ottima per ciascuna sottobanda si ottiene come

$$R_{1,ott} = R + \frac{1}{2}\log_2\left(\frac{\sigma_1^2}{\sqrt{\prod_{j=0}^{1}\sigma_j^2}}\right) = 3 + \frac{1}{2}\log_2\left(\frac{\frac{16}{17}}{\sqrt{\frac{16}{17}\cdot\frac{1}{17}}}\right) = 4$$

$$R_{2,ott} = R + \frac{1}{2}\log_2\left(\frac{\sigma_2^2}{\sqrt{\prod_{j=0}^{1}\sigma_j^2}}\right) = 3 + \frac{1}{2}\log_2\left(\frac{\frac{1}{17}}{\sqrt{\frac{16}{17}\cdot\frac{1}{17}}}\right) = 2 \qquad (7.23)$$

ed il guadagno di codifica che si ottiene è pari a

$$G_{SBC,max} = \frac{1}{2}\frac{\sum_{j=0}^{1}\sigma_j^2}{\sqrt{\prod_{j=0}^{1}\sigma_j^2}} = \frac{1}{2}\frac{1}{\sqrt{\frac{16}{17}\cdot\frac{1}{17}}} = \frac{17}{8} \qquad (7.24)$$

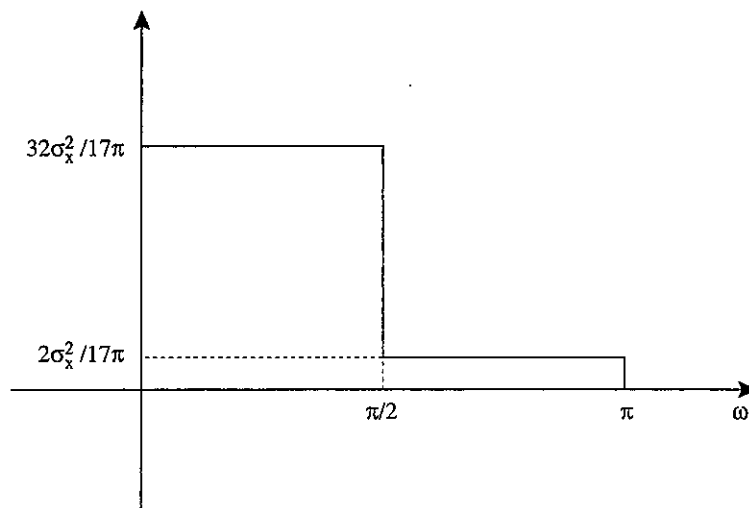



Fig. 7.5 - Esempio di distribuzione spettrale di potenza non uniforme.

Analogamente a quanto descritto a proposito della codifica predittiva, fissata l'entità del flusso numerico generato, il guadagno di codifica può essere

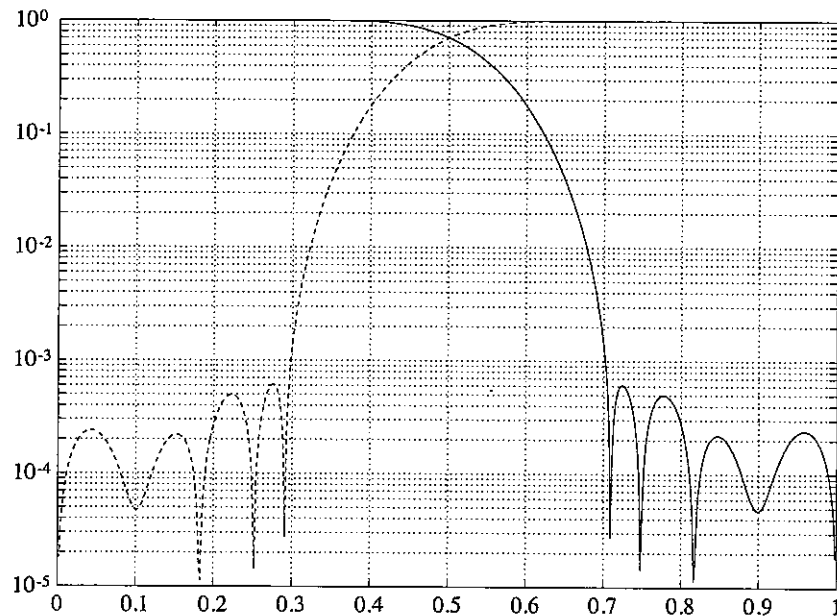



Fig. 7.6 - Funzioni di traferimento dei filtri utilizzati nella G.722.

sfruttato per migliorare la qualità del segnale, trasferendo bit dalle bande a dinamica inferiore verso quelle a dinamica maggiore. Viceversa, è possibile fissare la qualità del segnale prodotto e sfruttare il guadagno di codifica per comprimere il flusso numerico, eliminando bit nelle sottobanda a dinamica inferiore.

7.2.2 Raccomandazione ITU-T G.722

La raccomandazione ITU-T G.722 utilizza una codifica per sottobande di un segnale audio con 7 kHz di banda generando un flusso numerico di 64 kbit/s [ITU-T G.722].

Le bande utilizzate sono due, di 4 kHz ciascuna, ricavate tramite dei filtri QMF FIR a fase lineare con 24 coefficienti [Appendice D], il valore dei quali è riportato in tabella 7.1

In figura 7.6 sono riportate le relative funzioni di trasferimento. Date le uscite dei filtri

$$\begin{cases} x_A = \sum_{i=0}^{11} h_{2i}\, x_{in}(j-2i) \\ x_B = \sum_{i=0}^{11} h_{2i+1}\, x_{in}(j-2i-1) \end{cases} \qquad (7.25)$$

le componenti a frequenza inferiore e superiore del segnale si ottengono come

$$\begin{cases} x_L(n) = x_A + x_B \\ x_H(n) = x_A - x_B \end{cases} \qquad (7.26)$$

Coefficienti	Valore
h0, h23	0.000366211
h1, h22	- 0.00134277
h2, h21	- 0.00134277
h3, h20	0.00646973
h4, h19	0.00146484
h5, h18	- 0.0190430
h6, h17	0.00390625
h7, h16	0.0441895
h8, h15	- 0.0256348
h9, h14	- 0.0982666
h10, h13	0.116089
h11, h12	0.473145

Tab. 7.1 - Coefficienti dei filtri utilizzati nella G.722.

Le uscite dei due filtri sono ricampionate ad 8 kHz e ciascuna banda è codificata ADPCM. La banda superiore utilizza 2 bit per campione, producendo un flusso di 16 kbit/s. La banda inferiore può essere codificata secondo tre modalità, utilizzando 6, 5 o 4 bit per campione; i flussi numerici prodotti risultano di 48, 40 o 32 kbit/s. Nel caso di codifica su meno di bit inferiore a 6 bit per campione, fermo restando il flusso complessivo, si ricava un canale dati ausiliario di 8 o 16 kbit/s sfruttando i bit meno significativi del codice.

La codifica della banda inferiore su 6 bit avviene tramite un quantizzatore non uniforme su 60 livelli. Non si utilizza il numero massimo di livelli per un quantizzatore su 6 bit (=64), per evitare lunghe sequenze di zeri nel

flusso generato [ITU-T G.802]. La quantizzazione è adattativa e per l'aggiornamento del quanto si utilizzano solamente i quattro bit più significativi. In questo modo l'algoritmo non è influenzato dalla modalità di funzionamento.

Il predittore è simile a quanto descritto a proposito della raccomandazione G.721, con una struttura data dalla combinazione di un filtro a soli poli del 2° ordine ed uno a soli zeri del 6° ordine. I coefficienti di entrambi i filtri sono adattati tramite LMS con crosscorrelazione di polarità tra segnale ed errore.

7.2.3 Codifica per sottobande tramite modelli percettivi

La codifica di forma d'onda per sottobande precedentemente esposta tende alla ricostruzione fedele del segnale emesso dalla sorgente tramite una quantizzazione adattativa variabile per le differenti componenti spettrali. D'altra parte, grazie alla rappresentazione in frequenza fornita dalla codifica per sottobande, sarebbe possibile ottenere migliori rapporti di compressione sfruttando gli aspetti percettivi dell'apparato uditivo, eliminando componenti non rilevanti del segnale.

In particolare, la soglia di udibilità statica non è uniforme in frequenza, per cui è possibile sopprimere in fase di codifica quelle componenti spettrali la cui ampiezza è inferiore al valore della soglia per quella particolare frequenza. Inoltre, per quanto riguarda le componenti non mascherate e continuando a considerare esclusivamente la soglia di udibilità statica, la sua non uniformità permette di distribuire non uniformemente lo spettro del rumore di quantizzazione utilizzando quanti differenti per ciascuna sottobanda.

Il vantaggio maggiore, però, si ha considerando gli effetti di mascheramento. Come descritto nella presentazione dell'apparato uditivo, si ha una variazione locale e variabile nel tempo della soglia di udibilità in corrispondenza delle componenti spettrali del segnale. Questo permette di fissare per ciascuna sottobanda il livello ammissibile del rumore di quantizzazione, assegnando un opportuno numero di bit, in modo tale che questo si trovi al di sotto della soglia di udibilità istantanea e, quindi, non sia percepibile. Inoltre è possibile evitare totalmente la trasmissione di informazioni relative a sottobande che risultano totalmente mascherate.

Ciò richiede la realizzazione di un modello dell'apparato uditivo, detto modello psicoacustico o percettivo, in grado di fornire in tempo reale la soglia

di udibilità dinamica associata al particolare segnale in esame (fig. 7.7). Il calcolo della soglia di udibilità da parte del modello percettivo richiede innanzitutto la determinazione dello spettro del segnale. Data la necessità di ottenere un'elevata risoluzione, per il mappaggio in frequenza del segnale richiesto dal modello non possono essere utilizzate le uscite del banco di filtri (solitamente in numero limitato). Tipicamente, invece, questo è ottenuto tramite una FFT [Appendice E], eseguita, parallelamente al filtraggio. su di un blocco consistente di campioni del segnale d'ingresso.

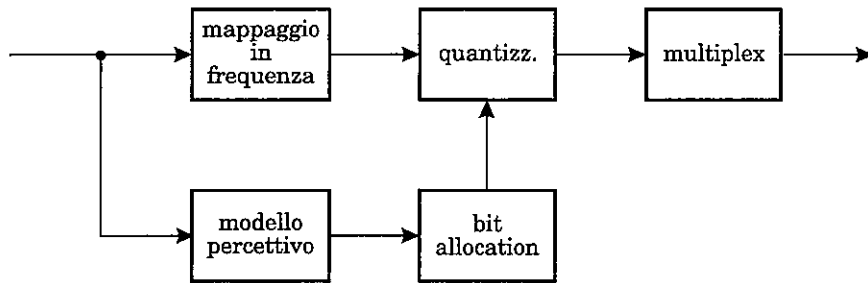



Fig. 7.7 - Codifica basata su modelli percettivi.

Lo spettro del segnale viene utilizzato per calcolare la soglia di udibilità istantanea. Nota la soglia di udibilità statica, il calcolo della soglia di udibilità istantanea richiede innanzitutto che vengano individuate le componenti mascheranti del segnale cercando i massimi relativi dello spettro del segnale (fig. 7.8). Per la differente capacità mascherante è poi opportuno determinare se tali componenti siano assimilabili a toni puri (segnali sinusoidali) o a rumore. Ciò si ricava controllando che lo spettro decada più o meno rapidamente da ciascun massimo relativo. Note le ampiezze delle componenti tonali e non tonali, è infine possibile determinare le caratteristiche di mascheramento e, quindi, la soglia di udibilità istantanea.

Nota la soglia di udibilità istantanea, viene calcolato per ciascuna sottobanda il rapporto segnale/maschera (Signal to Mask Ratio: SMR), come rapporto tra il livello di pressione sonora (massima componente spettrale) e minimo della soglia. In funzione di tale rapporto è possibile, infine, procedere alla definizione delle caratteristiche del quantizzatore per ciascuna sottobanda (bit allocation). Questa è una procedura iterativa che consiste nel non assegnare

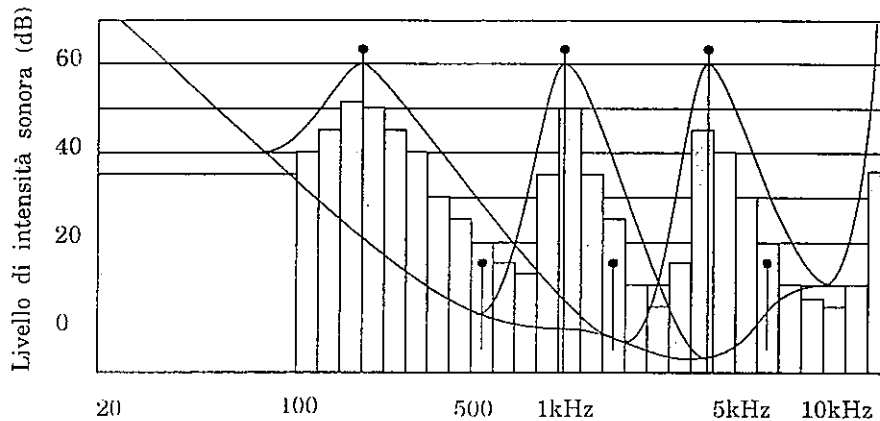



Fig. 7.8 - Rumore di quantizzazione permesso per sottobanda grazie al mascheramento con un rapporto maschera/rumore di 5 dB

bit in quelle bande dove l'informazione è posta al di sotto della soglia di udibilità istantanea e nell'aumentare progressivamente il numero di livelli di quantizzazione (migliorando, quindi, il rapporto segnale/rumore SNR) a partire da quelle sottobande dove risulta minimo il rapporto tra livello della maschera e quello del rumore di quantizzazione (Mask to quantizing Noise Ratio: MNR), definito come

$$MNR = SNR - SMR \qquad (7.27)$$

Fissato il flusso numerico e quindi il numero complessivo di bit utilizzabili per la codifica, con tale procedura si tenta di mantenere il rumore di quantizzazione almeno al di sotto di una certa soglia rispetto al livello della maschera (es.: 5 dB). Per le sottobande con segnale al di sotto della maschera, quindi, non viene trasmessa nessuna informazione. Per le altre, l'allocazione dei bit è ripetuta fino a che, per ciascuna finestra temporale (es.: 24 ms), non siano stati impegnati tutti i bit disponibili. La codifica che ne deriva risulta essere tendenzialmente a lunghezza variabile, anche se le principali applicazioni sono, invece, a bit rate costante.

7.2.4 Standard ISO/MPEG-1/Audio: Layer I

Un'implementazione di codifica per sottobande che tenga conto di modelli percettivi è stata proposta nel 1988 dall'ISO con standard MPEG-1 Audio [ISO92], approvato poi nel 1992. Questo standard prevede 3 differenti tecniche di codifica, delle quali la prima (Layer I) deriva dal MUSICAM (Masking-pattern adapted Universal Sub-band Integrated Coding And Multiplexing), uno standard proposto da CCETT, IRT, Philips e Matsushita.

Il segnale in ingresso (campionato a 32, 44.1 o 48 kHz e quantizzato su 16 bit) viene codificato in blocchi di 384 campioni. Il bit rate di uscita è fisso, ma selezionabile tra 14 differenti possibili valori che vanno da 32 a 448 kbit/s.

Il blocco d'ingresso è mappato in frequenza in 32 sottobande equispaziate di 750 Hz, con un campionamento di 48 kHz (fig. 7.9). Si nota come questa risoluzione è insufficiente per le bande critiche a frequenza più bassa, le cui ampiezze sono dell'ordine dei 100 Hz. La scomposizione in sottobande viene effettuata tramite un banco di 32 filtri FIR polifase a 512 coefficienti con cancellazione dell'aliasing [Appendice D], seguiti da una traslazione in banda base e da una decimazione con rapporto 32:1 (fig. 7.10). L'uscita del filtro è quindi esprimibile come:

$$s_i = \sum_{k=0}^{511} x(k) \cdot h(k) \cdot \cos\left[(2i+1)(k-16)\frac{\pi}{64} \right]$$
$$= \sum_{k=0}^{63} \cos\left[(2i+1)(k-16)\frac{\pi}{64} \right] \cdot \sum_{j=0}^{7} x(k+64\,j) \cdot h(k+64\,j) \qquad (7.28)$$

Ciascun blocco di 12 campioni consecutivi prodotti da ciascun filtro (384/32 campioni del segnale) è codificato tramite quantizzazione adattativa. Viene, infatti, trasmesso un fattore di scala (scalefactor), dato dal massimo valore assoluto delle ampiezze, che rappresenta la soglia di saturazione del blocco corrente. Codificando il fattore di scala su 6 bit ed utilizzando una risoluzione di 2 dB, la dinamica che ne deriva è di 96 dB. Il flusso necessario per la trasmissione di tale informazione è pari a 8 kbit/s.

Ricavato il fattore di scala, i campioni di ciascuna sottobanda sono poi normalizzati nell'intervallo [-1, 1]. Questa caratteristica, che fa pensare ad una rappresentazione in virgola mobile delle ampiezze, fa spesso vedere il fattore

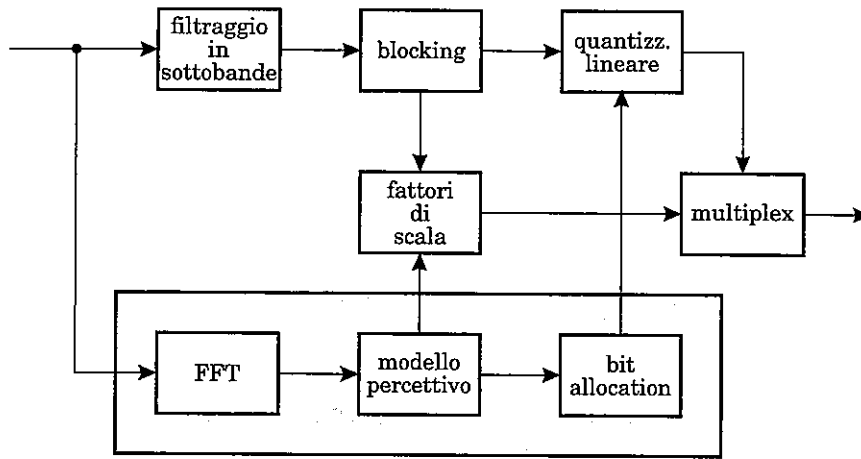



Fig. 7.9 - Codifica MPEG-1/Audio/Layer I.

di scala come esponente, dove la quantizzazione adattativa dei campioni rappresenta la mantissa (wordlength).

La codifica dei campioni avviene sfruttando il modello percettivo descritto nel paragrafo precedente (Modello I). Infatti, parallelamente al filtraggio, viene eseguita, per ciascuna finestra di 24 ms, una FFT su 1024 campioni del segnale d'ingresso per ottenere una stima del suo spettro con risoluzione (47 Hz) maggiore di quella permessa dal banco di filtri.

Noto lo spettro viene calcolata la soglia di udibilità istantanea ed il rapporto segnale/maschera per ciascuna sottobanda. La bit (o noise) allocation determina il numero di bit per sottobanda, con l'obiettivo di mantenere il rumore di quantizzazione almeno al di sotto di 5 dB del livello della maschera. Per le sottobande con segnale al di sopra della maschera, viene generata una codifica con una lunghezza variabile da 2 a 15 bit. Per le altre bande non viene trasmessa nessuna informazione. L'informazione su quale allocazione sia stata effettuata è codificata su 4 bit, con un flusso di 3.5 kbit/s.

La trama generata (fig. 7.11), prevede:

- un'intestazione contenente informazioni sulle caratteristiche del segnale d'ingresso (32 bit);
- un blocco per la rilevazione degli errori (16 bit);
- la codifica della bit allocation (32*4 bit);

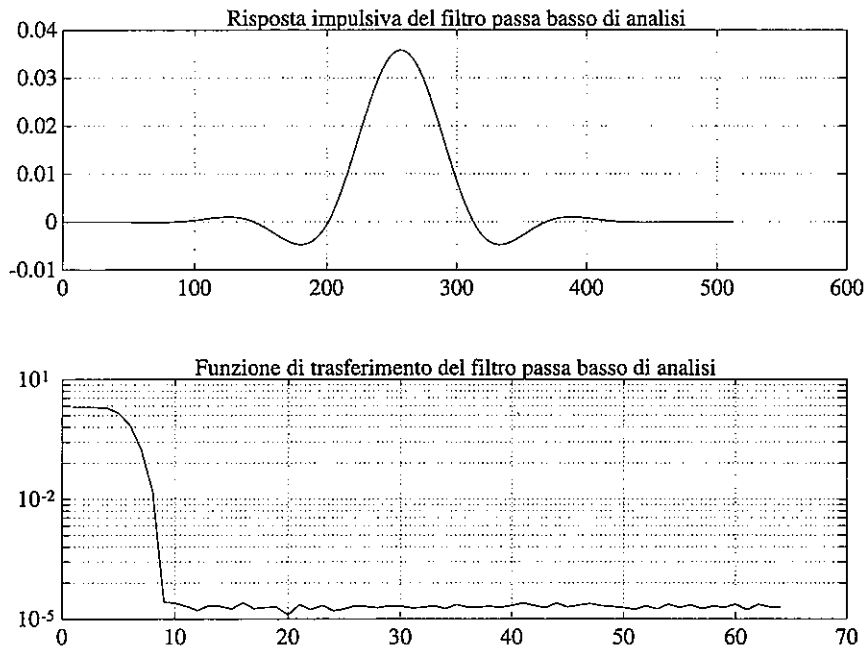



Fig. 7.10 - Caratteristiche dei filtri polifase.

- la codifica dei fattori di scala per le sottobande non eliminate ([0..31]*6 bit);
- la codifica dei campioni delle sottobande non eliminate ([0..31]*[2..15]*6 bit).

La decodifica è decisamente più semplice della codifica, al fine di ridurre il costo dei riproduttori, e si svolge le quattro fasi di decodifica della bit allocation, decodifica dei fattori di scala, riquantizzazione dei campioni e sintesi per sottobande.

In funzione delle informazioni ricevute sul numero "b" dei bit allocati in ciascuna sottobanda, si esegue la riquantizzazione dei campioni ricevuti "q" come

$$\hat{x} = \frac{2^b}{2^b - 1}\left(q + 2^{-b+1}\right) \tag{7.29}$$

Il risultato è un numero frazionale dal quale è possibile ottenere le uscite moltiplicandolo per il relativo fattore di scala "f". Il segnale di ciascuna sottobanda è, infine, inviato al banco di filtri di ricostruzione.

32	16	32×4	[0..31]×6	[0..31]×[2..15]×12
header	CRC	bit allocation	fattori di scala	campioni

Fig. 7.11 - Formato della trama per il Layer I della codifica MPEG-1/Audio.

È da notare come la struttura gli algoritmi che definiscono la quantizzazione delle uscite del banco dei filtri (modello psicoacustico ed algoritmo di bit allocation) siano del tutto trasparenti in decodifica, dato che nella trama trasmessa appaiono solamente il risultato della bit allocation, i fattori di scala ed i campioni delle sottobande. Di conseguenza è possibile utilizzare algoritmi più sofisticati in fase di codifica senza alterare il formato della trama trasmessa e, quindi, la struttura del decodificatore/riproduttore. Al termine della trama è prevista la possibilità di appendere dati ausiliari (ancillary data), utilizzabili per future estensioni dello standard.

Passando alle prestazioni, si nota come l'analisi in frequenza eseguita dal modello psicoacustico non abbia la stessa risoluzione temporale delle uscite dei filtri da codificare. Inoltre, nel caso in cui il blocco di campioni da codificare presenti variazioni di dinamica al proprio interno, il modello causerà la generazione di un rumore di quantizzazione in funzione della potenza media all'interno del blocco, che potrebbe risultare eccessivo, e quindi, percepibile, per le parti del segnale a dinamica inferiore (pre-eco) (fig. 7.12). Infine, dato che l'effetto di mascheramento si riduce al diminuire del livello, per porzioni di segnale con dinamica e conseguente mascheramento modesti, il flusso numerico (fisso) previsto potrebbe non essere sufficiente a garantire il voluto rapporto segnale rumore.

Lo standard MPEG, presentato per la codifica di un solo canale, permette, in realtà, la codifica di segnali stereo. Inoltre, esso è stato mantenuto anche nel- l'MPEG-2, che introduce una codifica multicanale 5+1 (anteriore: destro, centrale, sinistro; posteriore: destro, sinistro; effetti). Sono, inoltre, introdotte codifiche a basso bit rate con frequenze di campionamento di 16, 22.5 e 24 kHz [Bra95].

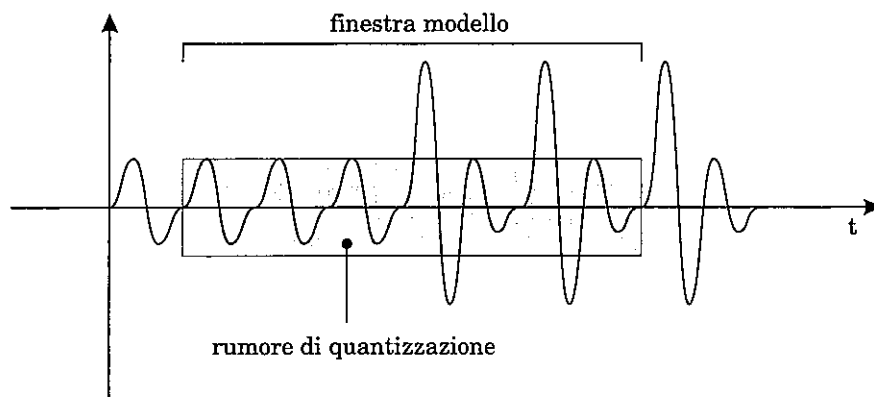



Fig. 7.12 - Pre-eco dovuto ad un transitorio di dinamica.

7.2.5 Codifica PASC

Dal punto di vista delle applicazioni, il Layer I di MPEG si caratterizza per una contenuta complessità computazionale, cosa che risulta particolarmente interessante nell'elettronica consumer. In questo modo, infatti, è possibile contenere il costo del codificatore/decodificatore e quindi del registratore/riproduttore.

Il primo esempio di utilizzo di tecniche percettive in sistemi consumer è stato la DCC (Digital Compact Cassette) della Philips [Wir91]. In questo caso, per la codifica del segnale è stato utilizzato lo standard PASC (Precision Adaptive Sub-band Coding), compatibile con l'MPEG Layer I. Il PASC adottato nella DCC, infatti, è praticamente l'MPEG Layer I con flusso numerico fissato a 192 kbit/s per canale. Il fattore di compressione complessivamente ottenuto rispetto al PCM è pari a quattro.

La necessità di comprimere il segnale nasce dall'esigenza di mantenere una compatibilità meccanica tra DCC e cassetta analogica, sia in termini di velocità di trascinamento del nastro, che nell'utilizzo di una testina di lettura/scrittura fissa. In tal modo, pur utilizzando in parallelo otto tracce, non è possibile ottenere una velocità di lettura/scrittura sufficiente a sostenere il flusso di 1.5 Mbit/s prodotto da segnali con frequenze di campionamento dell'ordine dei 40 kHz e 16 bit per campione (es.: 44.1 kHz/16 bit per il Compact Disc). Il problema di sostenere senza compressione un flusso numerico paragonabile a quello del CD, aveva portato nel DAT (Digital Audio

Tape) della Sony, all'adozione di un sistema meccanico di lettura/scrittura basato su di una scansione elicoidale del nastro tramite una testina rotante, analogamente a quanto già adottato nei sistemi di videoregistrazione.

7.2.6 Standard ISO/MPEG-1/Audio: Layer II

Sempre nell'MPEG Audio, si trova, come Layer II, un algoritmo che, pur derivando anch'esso dal MUSICAM, è in grado di ottenere rapporti di compressione più elevati del Layer I tramite una codifica di sorgente dei dati prodotti e lo sfruttamento di fenomeni di mascheramento temporale. Queste migliori caratteristiche si traducono, però, in un conseguente incremento della complessità del codificatore/decodificatore (fig. 7.13). Il bit rate di uscita è sempre fisso, ma selezionabile tra 14 differenti possibili valori che vanno da 32 a 384 kbit/s.

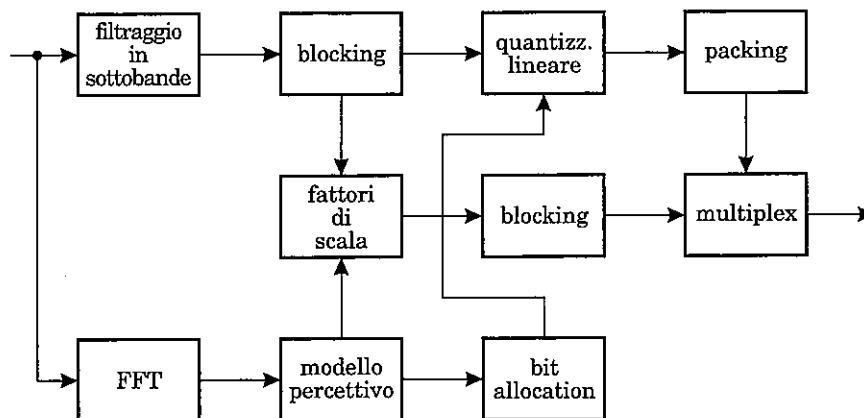



Fig. 7.13 - Codi fica MPEG-1/Audio Layer II.

Pur utilizzando lo stesso modello psicoacustico del Layer I, nella codifica si considerano 3 sottoblocchi consecutivi di 384 campioni, con 1152 campioni complessivi per blocco. L'obiettivo del raggruppamento di tre sottoblocchi del segnale, è quello di condividere gli stessi codici che identificano la quantizzazione implementata, sfruttando il mascheramento temporale dell'apparato uditivo. Infatti, in condizioni stazionarie, è possibile mantenere la stessa bit allocation per i 3 blocchi, mentre in presenza di transitori, è necessario trasmettere due o tutti e tre i codici di bit allocation. Per decidere se è possibile

raggruppare i fattori di scala, vengono definite due "classi", date delle differenze tra i fattori di scala del primo e secondo sottoblocco e tra il secondo ed il terzo sottoblocco. In funzione del valore di tali classi, è possibile trasmettere non tutti i tre fattori di scala, ma solamente alcune loro combinazioni. Ad esempio, è possibile trasmettere solo il primo, il primo fattore di scala, solo il secondo, il massimo dei tre, ecc. Ovviamente è necessario trasmettere l'informazione su quali raggruppamenti sono stati effettuati tramite una Scale Factor Selection Information (SCFSI), codificata su 2 bit per ciascun blocco. Tale informazione è inserita nella trama tra l'allocazione ed i fattori di scala (fig. 7.14).

					[0..11]×[0..31]×[5..10]
32	16	[0..31]×[2..4]	[0..31]×2	[0..31]×[1..3]×6	[0..11]×[0..31]×[0..2]×[2..]
header	CRC	bit allocation	scfsi	fattori di scala	campioni

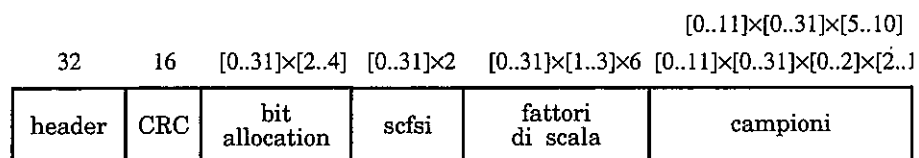



Fig. 7.14 - Formato della trama per il Layer II della codifica MPEG-1/Audio.

Un'ulteriore compressione è ottenibile nella codifica delle ampiezze dei campioni. Infatti, nel caso si utilizzino quantizzazioni su pochi bit (3, 5 o 9 livelli), vengono raggruppate le uscite di ciascuna sottobanda in blocchi di 3 campioni consecutivi, chiamati granuli, codificati tramite un unico codice.

Campo di applicazione del Layer II è nella codifica di segnali per trasmissioni broadcast (Digital Audio Broadcasting: DAB) a 128 kbit/s per canale o con 4 canali a 192 kbit/s su di un unico canale DSR (Digital Satellite Radio) non compresso.

7.3 CODIFICA PER TRASFORMATE

7.3.1 Generalità sulla codifica per trasformate

Si consideri una trasformazione espressa tramite una matrice A di dimensioni N×N che, a partire dal vettore di N campioni $\mathbf{x} = \{x(0), x(1), ..., x(N-1)\}^T$ fornisca un vettore di coefficienti y di pari lunghezza

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} \tag{7.30}$$

Si indichi con **B** la trasformazione inversa che, a partire dal vettore di coefficienti **y** riproduca un vettore $\hat{\mathbf{x}}$, generalmente diverso da **x** a causa delle approssimazioni introdotte nella codifica di **y**. Si vuole individuare la coppia di trasformazioni **A** e **B** tali che la quantizzazione e codifica dei coefficienti risulti vantaggiosa secondo qualche criterio. In particolare:

- qualora la trasformazione riuscisse a concentrare l'energia su di un numero ridotto di coefficienti con un andamento dell'ampiezza degli stessi risultasse decrescente (con lunghe sequenze di coefficienti nulli), sarebbe possibile ridurre il numero medio di bit per campione sia applicando tecniche di compressione del tipo run-length, sia con codifica di Huffman dei coefficienti della trasformata;
- la trasformazione dovrebbe risultare ottima secondo qualche criterio, ad esempio tale da minimizzare l'errore quadratico medio di ricostruzione

$$\frac{1}{N} E\left\{ \sum_{k=1}^{N} (x_k - \hat{x}_k)^2 \right\} \tag{7.31}$$

Prima di affrontare il problema della determinazione della trasformata ottima, è necessario richiamare alcune proprietà di analisi delle matrici. Per una sequenza mono-dimensionale **x**, una matrice **A** rappresenta una trasformazione in un vettore **y**, i cui elementi si ottengono come

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} \;\Rightarrow\; y(k) = \sum_{n=0}^{N-1} a(k,n)\, x(n)\,, \quad 0 \le k \le N-1 \tag{7.32}$$

La matrice **A** è detta ortogonale se la sua inversa coincide con la sua trasposta

$$\mathbf{A}^{-1} = \mathbf{A}^{T} \;\Rightarrow\; \mathbf{A}\,\mathbf{A}^{T} = \mathbf{A}^{T}\mathbf{A} = \mathbf{I} \tag{7.33}$$

Nel caso in cui la matrice **A** sia complessa, è detta unitaria se la sua inversa coincide con la trasposta coniugata

$$\mathbf{A}^{-1} = \mathbf{A}^{*T} \tag{7.34}$$

Per una matrice unitaria, la trasformazione inversa avviene come

$$\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{y} = \mathbf{A}^{*T}\,\mathbf{y} \;\Rightarrow\; x(n) = \sum_{k=0}^{N-1} a^*(k, n)\, y(k)\,, \quad 0 \le n \le N-1 \tag{7.35}$$

Una matrice ortogonale reale è anche unitaria, ma non è vero il viceversa. Inoltre, come conseguenza diretta della definizione, si ha che le righe o le colonne di una matrice unitaria formano una base ortogonale nello spazio ad N dimensioni. Le colonne di $\mathbf{A}^T$, cioè i vettori $\mathbf{a}_k^* = a^*(k, n)$, $0 \le n \le N-1$ vengono chiamati vettori base di A.

Una matrice **A** è detta simmetrica se coincide con la sua trasposta ($\mathbf{A} = \mathbf{A}^T$) ed Hermitiana se coincide con la trasposta coniugata ($\mathbf{A} = \mathbf{A}^T$). Per una matrice Hermitiana tutti gli autovalori (radici dell'equazione $|\,\mathbf{A} - \lambda_k\,\mathbf{I}\,| = 0$) sono reali. Inoltre, esiste sempre la matrice unitaria degli autovettori $\mathbf{\Phi}$ (automatrice di A) tale che

$$\mathbf{\Phi}^T\,\mathbf{A}\,\mathbf{\Phi} = \Lambda \tag{7.36}$$

dove Λ è la matrice diagonale formata dagli autovalori di **A**.

Alcune proprietà delle trasformate unitarie sono particolarmente importanti per quanto riguarda la codifica. La prima è la conservazione dell'energia. Infatti

$$\|\,\mathbf{y}\,\|_2 = \sum_{k=0}^{N-1} |\,y(k)\,|^2 = \mathbf{y}^T\,\mathbf{y} = \mathbf{x}^T\,\mathbf{A}^T\;\mathbf{A}\,\mathbf{x} = \mathbf{x}^T\,\mathbf{x} = \sum_{k=0}^{N-1} |\,x(k)\,|^2 = \|\,\mathbf{x}\,\|_2 \tag{7.37}$$

Ciò significa che, interpretando i campioni del segnale come coordinate di uno spazio ad N dimensioni, ogni trasformazione unitaria è semplicemente una rotazione del vettore x, fermo restando il suo modulo.

Questa proprietà apre la strada alla ricerca di trasformate unitarie che compattino l'energia del segnale in un numero ridotto di coefficienti. Infatti, poiché l'energia viene conservata, se essa viene compattata, molti coefficienti conterrebbero un'energia molto piccola e, quindi, risulterebbero trascurabili. Condizione necessaria perché ciò avvenga è che i coefficienti risultino essere scorrelati: in tal modo è possibile che alcuni se ne annullino, mentre altri risultano essere diversi da zero.

Se μ_x e $\mathbf{R}_x$ indicano la media e la covarianza del vettore **x**, allora le quantità corrispondenti per il vettore trasformato **y** sono date da

$$\mu_y = E[\mathbf{y}] = E[\mathbf{A}\,\mathbf{x}] = \mathbf{A}\,E[\mathbf{x}] = \mathbf{A}\,\mu_x$$

$$\mathbf{R}_y = E\left[(\mathbf{y}-\mu_y)(\mathbf{y}-\mu_y)^T\right] = \mathbf{A}\left\{E\left[(\mathbf{x}-\mu_x)(\mathbf{x}-\mu_x)^T\right]\right\}\mathbf{A}^T = \mathbf{A}\,\mathbf{R}_x\,\mathbf{A}^T \tag{7.38}$$

La varianza dei coefficienti è data dagli elementi della diagonale di $\mathbf{R}_y$, cioè

$$\sigma_y^2(k) = [\,\mathbf{R}_y\,]_{k,k} = \left[\,\mathbf{A}\,\mathbf{R}_x\,\mathbf{A}^T\,\right]_{k,k} \tag{7.39}$$

Dato che **A** è unitaria

$$\begin{cases} \sum_{k=0}^{N-1} |\,\mu(k)\,|^2 = \mu_y^T\,\mu_y = \mu_x^T\,\mathbf{A}^T\,\mathbf{A}\,\mu_x = \sum_{n=0}^{N-1} |\,\mu_x(n)\,|^2 \\ \sum_{k=0}^{N-1} \sigma_y^2(k)\,| = \mathrm{Tr}\left[\,\mathbf{A}\,\mathbf{R}_x\,\mathbf{A}^T\,\right] = \mathrm{Tr}\,[\,\mathbf{R}_x\,] = \sum_{n=0}^{N-1} |\,\sigma_x^2(n)\,|^2 \end{cases}$$

$$\sum_{k=0}^{N-1} E\left[\,|\,y(k)\,|^2\,\right] = \sum_{n=0}^{N-1} E\left[\,|\,x(n)\,|^2\,\right] \tag{7.40}$$

Se si considera come matrice di trasformazione l'automatrice Φ della matrice di autocovarianza **R**

$$\begin{cases} \mathbf{X} = \Phi^T\,\mathbf{x} \\ \mathbf{x} = \Phi\,\mathbf{X} \end{cases} \tag{7.41}$$

si ottiene la trasformata di Karhunen-Loeve (KLT). Per gli elementi **X** della sequenza trasformata, risulta

$$E\left[\,\mathbf{X}\,\mathbf{X}^T\,\right] = \Phi^T\,E\left[\,\mathbf{X}\,\mathbf{X}^T\,\right]\Phi = \Phi^T\,\mathbf{R}\,\Phi = \Lambda$$

$$E\left[\,X(k)\,X(l)^*\,\right] = \lambda_k\,\delta(k-l) \tag{7.42}$$

cioè essi risultano essere ortogonali e scorrelati. In questo modo si soddisfa il desiderato criterio di indipendenza dei coefficienti della trasformata.

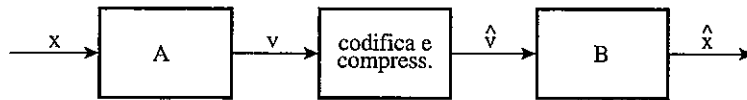



Fig. 7.15 - Codifica per trasformate.

Si consideri, ora, il problema della minimizzazione dell'errore quadratico medio. Si vuole verificare che l'insieme di vettori base Φ_i , i=0,...,N−1 minimizza l'MSE di una rappresentazione troncata, utilizzante solo i primi D coefficienti (D < N) di una coppia di trasformazioni **A-B**. L'importanza dal punto della codifica di tale approccio è evidente, dato che la mancata trasmissione dei coefficienti irrilevanti comporterebbe una compressione del flusso numerico prodotto (fig. 7.15). La rappresentazione troncata $\hat{\mathbf{x}}$ di **x** è data da

$$\hat{x} = \sum_{i=0}^{D-1} X_i \Phi_i \qquad (7.43)$$

ed il corrispondente MSE si ottiene come

$$\varepsilon_D = \frac{1}{N} E\left[\sum_{n=0}^{N-1} | x(n) - \hat{x}(n) |^2\right] = \frac{1}{N} \mathrm{Tr}\left\{E\left[(x-\hat{x})(x-\hat{x})^T\right]\right\} \qquad (7.44)$$

Introducendo una matrice identità $\mathbf{I}_D$ di dimensione D, l'MSE può essere riscritto come

$$\varepsilon_D = \frac{1}{N} \mathrm{Tr}\left[(\mathbf{I} - \mathbf{B}\,\mathbf{I}_D\,\mathbf{A})\,\mathbf{R}\,(\mathbf{I} - \mathbf{B}\,\mathbf{I}_D\,\mathbf{A})^T\right] \qquad (7.45)$$

La minimizzazione dell'MSE è ottenibile differenziando questa espressione rispetto ad **A** ed uguagliando a zero

$$\mathbf{I}_D\,\mathbf{B}^T (\mathbf{I} - \mathbf{B}\,\mathbf{I}_D\,\mathbf{A})\,\mathbf{R} = 0 \qquad (7.46)$$

da cui si ottiene l'errore

$$\varepsilon_D = \frac{1}{N} \mathrm{Tr}\left[(\mathbf{I} - \mathbf{B}\,\mathbf{I}_D\,\mathbf{A})\,\mathbf{R} \right]$$

$$\mathbf{I}_D\,\mathbf{B}^T = \mathbf{I}_D\,\mathbf{B}^T\,\mathbf{B}\,\mathbf{I}_D\,\mathbf{A} \tag{7.47}$$

Imponendo che l'errore si annulli per D = N

$$\mathbf{I} - \mathbf{B}\,\mathbf{A} = 0 \;\rightarrow\; \mathbf{B} = \mathbf{A}^{-1}$$

$$\mathbf{I}_D\,\mathbf{B}^T\,\mathbf{B} = \mathbf{I}_D\,\mathbf{B}^T\,\mathbf{B}\,\mathbf{I}_D \tag{7.48}$$

Affinché ciò sia vero per qualsiasi valore di D, la matrice **B** deve essere unitaria e, di conseguenza, risulterà unitaria anche la matrice **A** con $\mathbf{B} = \mathbf{A}^T$.

Per quanto riguarda l'errore

$$\varepsilon_D = \frac{1}{N} \mathrm{Tr}\left[(\mathbf{I} - \mathbf{A}^T\,\mathbf{I}_D\,\mathbf{A})\,\mathbf{R} \right] = \frac{1}{N} \mathrm{Tr}\left[\mathbf{R} - \mathbf{I}_D\,\mathbf{A}\,\mathbf{R}\,\mathbf{A}^T \right] \tag{7.49}$$

Dato che **R** è fissa, è necessario massimizzare la quantità

$$\hat{\varepsilon}_D = \mathrm{Tr}\left[\mathbf{I}_D\,\mathbf{A}\,\mathbf{R}\,\mathbf{A}^T \right] = \sum_{k=0}^{D-1} \mathbf{a}_k^T\,\mathbf{R}\,\mathbf{a}_k \tag{7.50}$$

dove $\mathbf{a}_k^T$ è la k-esima riga di **A** e, essendo **A** unitaria, $\mathbf{a}_k^T\,\mathbf{a}_k = 1$. A tal fine si considera il lagrangiano

$$\hat{\varepsilon}_D = \sum_{k=0}^{D-1} \mathbf{a}_k^T\,\mathbf{R}\,\mathbf{a}_k + \sum_{k=0}^{D-1} \lambda_k\,(1 - \mathbf{a}_k^T\,\mathbf{a}_k) \tag{7.51}$$

differenziando il quale rispetto ad $\mathbf{a}_j$ si ottiene la soluzione

$$\mathbf{R}\,\mathbf{a}_j = \lambda_j\,\mathbf{a}_j \tag{7.52}$$

che porta all'errore

$$\hat{\varepsilon}_D = \sum_{k=0}^{D-1} \lambda_k \tag{7.53}$$

Questo errore è massimo se $\mathbf{a}_j$ corrisponde al maggiore autovalore di **R**. Per massimizzare l'errore per qualsiasi valore di D, è necessario ordinare gli autovalori in ordine decrescente

$$\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{N-1} \tag{7.54}$$

Le righe di **A** sono quindi gli autovettori di **R** e, quindi, la trasformazione A porta di nuovo alla KLT. L'MSE dovuto al troncamento è dato da

$$\varepsilon_D = \sum_{i=D}^{N-1} \lambda_i \tag{7.55}$$

Una conseguenza importante di questa proprietà è che la KLT, tra tutte le trasformazioni unitarie, riesce più di altre a concentrare la maggior parte dell'energia in un numero D ≤ N di coefficienti. Infatti, se si considera la serie ordinata delle varianze dei coefficienti ottenuti da una trasformazione **A**

$$\sigma_k^2 = E\left[\, |X(k)|^2 \right], \quad \sigma_0^2 \geq \ldots \geq \sigma_k^2 \geq \ldots \geq \sigma_{N-1}^2 \tag{7.56}$$

e se ne considera una loro somma parziale

$$S_m(\mathbf{A}) = \sum_{k=0}^{m-1} \sigma_k^2, \quad m \in [1, N] \tag{7.57}$$

risulta

$$S_D(\mathbf{A}) = \sum_{k=0}^{D-1} (\mathbf{ARA^T})_{k,k} = \mathrm{Tr}\,(\mathbf{I}_D\, \mathbf{A^T}\, \mathbf{R}\, \mathbf{A}) = \hat{\varepsilon}_D \tag{7.58}$$

che essere massimizzato quando **A** è la trasformata KL. Poiché $\sigma_k^2 = \lambda_k$ quando $\mathbf{A} = \Phi^{\mathbf{T}}$, si ha

$$\sum_{k=0}^{D-1} \lambda_k \geq \sum_{k=0}^{D-1} \sigma_k^2, \quad 1 \leq D \leq N \tag{7.59}$$

e quindi

$$S_D(\Phi T) \geq S_D(A) \tag{7.60}$$

In tal modo, eliminando i coefficienti di ampiezza trascurabile, il vettore **x** può essere rappresentato da un numero ridotto di elementi, eseguendo la voluta compressione del segnale.

Nonostante le buone proprietà della KLT, risulta evidente che occorre trovare una trasformata più agevole, che non dipenda dai dati in ingresso e che sia di più semplice implementazione. Infatti le funzioni base dipendono dalla matrice di autocovarianza **R** e quindi dal segnale e non possono essere predeterminate a meno di pochi casi particolari nei quali sono disponibili soluzioni analitiche.

L'approssimazione migliore della KLT con coefficienti indipendenti dai dati è data dalla trasformata coseno di Fourier (FCT) e, nel discreto, dalla Discrete Cosine Transform (DCT) [appendice E] (fig. 7.16). A causa dei suoi legami con l'analisi in frequenza, nella codifica audio si preferisce usare questa trasformata anche perché legata agli aspetti psicofisici dell'udito, pur se lievemente meno efficiente dal punto di vista della compressione.

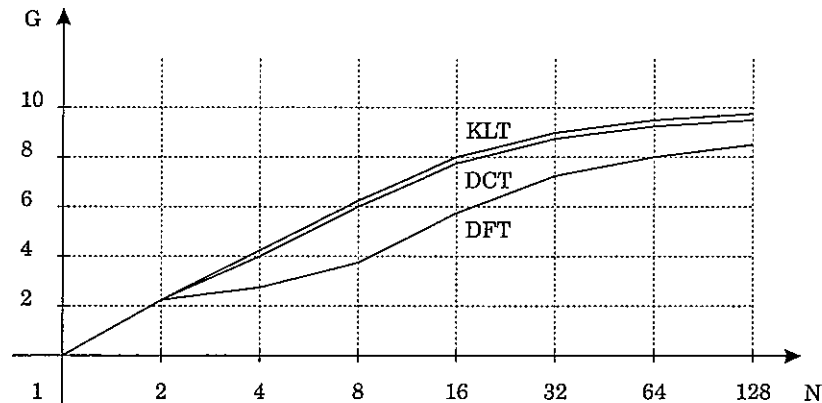



Fig. 7.16 - Guadagno rispetto al PCM di codifiche per trasformate.

Per la definizione della trasformata coseno, si consideri un segnale x(t) definito solo per $t \geq 0$ e si costruisca una funzione y(t) data da

$$y(t) = \begin{cases} x(t) & t \geq 0 \\ x(-t) & t \leq 0 \end{cases} \tag{7.61}$$

La trasformata di Fourier [Appendice E] di questa nuova funzione si ottiene come

$$F[y(t)] = \sqrt{\frac{1}{2\pi}} \left\{ \int_0^{\infty} x(t)\, e^{-j\omega t}\, dt + \int_{-\infty}^{0} x(-t)\, e^{j\omega t}\, dt \right\}$$

$$= \sqrt{\frac{1}{2\pi}} \int_0^{\infty} x(t) \left[e^{-j\omega t} + e^{j\omega t} \right] dt \;=\; \sqrt{\frac{1}{2\pi}} \int_0^{\infty} x(t) \cos(\omega t)\, dt \qquad (7.62)$$

dove l'ultimo membro, per definizione, rappresenta la trasformata coseno di Fourier di x(t).

In tal modo, la codifica per trasformate avviene calcolando dapprima la trasformata coseno della serie dei campioni e poi quantizzando i coefficienti ottenuti (Adaptive Transform Coding: ATC).

Nel rendere adattativa la quantizzazione, è possibile sfruttare un modello percettivo. Inoltre, data l'elevata probabilità che i coefficienti a frequenza maggiore si annullino dopo la quantizzazione, l'efficienza di compressione può essere ulteriormente aumentata tramite una codifica entropica dei coefficienti ottenuti.

7.3.2 Codifica ASPEC

La codifica ASPEC (Adaptive Spectral Perceptual Entropy Coding) è una codifica per trasformate proposta da AT&T, CNET, Fraunhofer Institute /Erlangen University e TCE.

La trasformata utilizzata è la DCT. Le finestre di campioni sulle quali viene eseguita la trasformazione si sovrappongono per metà della lunghezza del blocco al fine di ridurre eventuali discontinuità nel passaggio da un blocco al successivo. Dato che, come già visto, una variazione della dinamica del segnale sugli ultimi campioni della finestra si ripercuoterebbero sull'intero blocco, ciò provocherebbe in sede di decodifica un indesiderabile pre-eco. A tal fine la lunghezza della finestra stessa varia dinamicamente in modo di adattarsi alle caratteristiche del segnale. In particolare la lunghezza si riduce da 1024 a 256 campioni in corrispondenza di un gradino di dinamica. Questa variazione delle dimensioni della finestra è, ovviamente, segnalato nel flusso prodotto.

A causa della sovrapposizione tra finestre, blocchi lunghi e corti generano rispettivamente blocchi di 512 e 128 coefficienti. Tali coefficienti vengono codificati innanzitutto scalandoli tramite un opportuno fattore di scala. Al fine di ridurre il flusso richiesto alla trasmissione di tali fattori di scala, per ogni 1024 campioni, nel caso di blocchi lunghi i coefficienti vengono raggruppati in 20 bande, mentre nel caso di blocchi corti si utilizza una serie di 4 gruppi di 12 bande. La trasmissione di un fattore di scala per ciascuna banda genera un flusso relativo di 8 kbit/s.

La successiva codifica dei coefficienti utilizza un modello percettivo simile a quello utilizzato nel MUSICAM. Il calcolo della soglia istantanea tiene conto sia degli effetti di mascheramento all'interno della banda che delle interazioni tra bande adiacenti. Nota la maschera si procede alla bit allocation. L'informazione relativa a quale allocazione sia stata prescelta viene trasmessa con una banda di circa 2 kbit/s.

I coefficienti vengono, infine codificati tramite una codifica di Huffman, con lunghezza variabile da zero a 19 bit.

7.3.3 Standard ISO/MPEG-1/Audio: Layer III

Un algoritmo che combina le caratteristiche del MUSICAM e dell'ASPEC è utilizzato nel Layer III della codifica ISO/MPEG-1/Audio [ISO92].

Il segnale è suddiviso in blocchi di 1152 campioni. Il bit rate di uscita è variabile in tempo reale tra 14 differenti possibili valori che vanno da 32 a 320 kbit/s. Nel caso di codifica a bit rate costante, è possibile considerare una bufferizzazione (bit reservoir) che permette di spostare bit di blocchi più esigenti in termini di flusso su quelli che lo sono meno.

La principale differenza nella codifica tra i primi due livelli dell'MPEG ed il terzo è che, in questo caso, alle uscite del banco dei filtri viene applicata una trasformata, rendendo il codificatore ibrido (per sottobande e trasformate) (fig. 7.17). Altre differenze si hanno dall'adozione di una quantizzazione non uniforme, da una segmentazione adattativa e da una codifica entropica delle uscite.

La trasformata, che è una trasformata coseno modificata (MDCT), viene applicata alle uscite del banco di filtri per aumentare la risoluzione dell'analisi in frequenza. Infatti, nel Layer I si era visto come per le frequenze inferiori, la

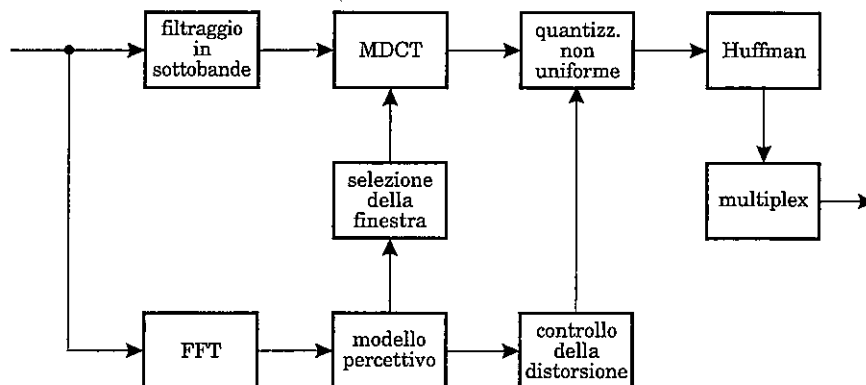



Fig. 7.17 - Codifica MPEG-1/Audio/Layer III.

larghezza di banda dei filtri (750 Hz) era eccessiva se paragonata a quella delle bande critiche (100 Hz) (vedi tab. 1.1). La trasformata viene eseguita su blocchi di dimensioni variabili da 12 (considerando separatamente ciascun sotto-blocco) a 18 campioni (considerando l'intero blocco scomposto in due parti). Questo è fatto per tenere in conto fenomeni di pre-eco dovuti a variazioni di dinamica del segnale. Considerando una MDCT su 18 campioni, si ottiene una risoluzione in frequenza di circa 40 Hz, sufficiente per risolvere le bande critiche. Ovviamente, la risoluzione temporale viene peggiorata.

Come detto, la dimensione della finestra di analisi è resa variabile in funzione delle caratteristiche del segnale, classificato in quattro differenti stati (normal, start, stop, short). La rilevazione di un transitorio avviene in funzione dello scostamento del risultato della bit allocation dal valor medio, con relativo accesso al bit reservoir. Questo approccio si giustifica per il fatto che nei transitori sono presenti sia campioni con ampiezze elevate (che portano ad innalzare la soglia di saturazione del quantizzatore e quindi il fattore di scala), sia porzioni di segnale con ampiezze modeste (che richiedono quanti di ampiezza ridotta e quindi elevato numero di livelli): questo complessivamente provoca un brusco aumento di bit richiesti nella codifica.

Il modello psicoacustico utilizzato per il Layer III è basato su FFT, ma si adotta una finestra variabile di 1024 o 256 campioni, in funzione della stazionarietà del segnale (Modello II). Inoltre, calcolato lo spettro in modulo e fase, tali valori vengono anche stimati da quelli ottenuti nei precedenti due blocchi. Per non appesantire eccessivamente gli aspetti computazionali, tale

predizione può essere limitata ad una banda inferiore ai 7Khz o, al massimo, di 3kHz. Lo spettro attuale e la sua stima sono utilizzati per fissare una soglia di mascheramento che tenga in conto fenomeni di pre-eco.

Passando alla codifica dei coefficienti della MDCT, la quantizzazione adottata è logaritmica. Per controllare che la soglia di mascheramento calcolata dal modello sia rispettata, questa codifica (che è a perdita) è eseguita in maniera iterativa in due loop annidati, uno di calcolo ed uno di verifica.

La codifica dei fattori di scala è simile a quanto fatto per il Layer II. La quantizzazione è relativa a coppie di frequenze (granuli) ed è possibile trasmettere un solo fattore di scala per due granuli adiacenti. La codifica è a lunghezza variabile tramite una codifica statica di Huffman.

Un possibile campo di applicazione del Layer III è nella trasmissione di segnali a banda audio (20 kHz) su canali a 64 kbit/s (es.: ISDN).

7.3.4 Codifica ATRAC

La codifica ATRAC (Adaptive Transform Acoustic Coding) è la codifica utilizzata per il MiniDisc della Sony.

Anche questo è un codificatore ibrido, che utilizza una trasformata in cascata ad un banco di filtri (fig. 7.18). Per garantire una risoluzione paragonabile a quella delle bande critiche (migliore alle basse frequenze), il banco di filtri è costituito da filtri QMF [Appendice D], con bande di 0-5.5 kHz, 5.5-11 kHz e 11-22.05 kHz. All'uscita dei filtri è applicata una MDCT. Infine, coefficienti relativi a bande adiacenti sono raggruppati non uniformemente in blocchi chiamati Block Floating Unit (BFU), dando alle basse frequenze una risoluzione migliore delle alte.

Le finestre temporali utilizzate per la MDCT, che presentano overlap tra blocchi adiacenti (al fine di ottenere una buona risoluzione), hanno ampiezze che variano in maniera adattativa in funzione delle caratteristiche del segnale. Si hanno due modalità di funzionamento. La prima (long mode) ha una risoluzione temporale di 11.6 ms (512 campioni). La seconda modalità (short mode) ha una risoluzione di 1.45 ms (64 campioni) alle frequenze superiori e di 2.9 ms (128 campioni) alle inferiori. Alla buona risoluzione temporale alle bande superiori corrisponde una bassa risoluzione in frequenza e viceversa per le bande inferiori, coerentemente con gli aspetti percettivi. Il passaggio allo

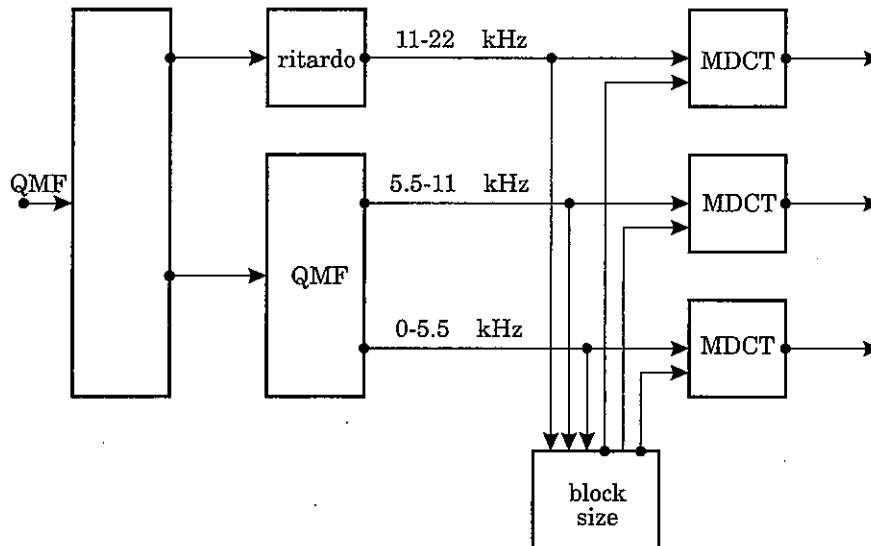



Fig. 7.18 - Struttura del codificatore ATRAC.

short mode avviene solo per transitori dati da incrementi di dinamica (attack), mentre nei transitori dovuti a riduzioni di dinamica (decay) si continua ad operare in long mode. Questo è giustificato dal fatto che il forward masking è molto più efficiente del backward masking.

La trama prodotta contiene:

- MDCT block size (long/short)
- word length per ciascuna BFU
- scale factor per ciascuna BFU
- coefficienti spettrali quantizzati

Per quanto riguarda la bit allocation, lo standard non specifica nessun algoritmo, dato che il decodificatore ne è indipendente e si permette, in questo modo, un'evoluzione delle tecniche adottate. Comunque, in [Tsu92] viene descritta una possibile implementazione avente come obiettivo una bassa complessità computazionale. Questa è basata su una prima assegnazione di bit fissa, che privilegia le frequenze inferiori ed una seconda che è funzione del logaritmo dello spettro all'interno di ciascuna BFU.

Con un ingresso campionato a 44.1 su 16 bit, il rapporto di compressione è di 5:1 ed il bit rate di uscita è di 140 kbit/s per canale.

Appendice A

PROCESSI AUTOREGRESSIVI

A.1 SISTEMI LINEARI TEMPO INVARIANTI A TEMPO DISCRETO

Un sistema discreto è individuato dalla trasformazione T che lega l'ingresso x con l'uscita y

$$y(n) = T\{ x(n) \} \tag{A.1}$$

Affinché un sistema discreto sia lineare deve risultare che la sua uscita ad una combinazione lineare di ingressi sia la stessa combinazione lineare alle singole uscite, cioè

$$T\{a\, x_1(n) + b\, x_2(n)\} = a\, T\{x_1(n)\} + b\, T\{x_2(n)\} \tag{A.2}$$

dove a e b sono delle costanti. Invece, un sistema è detto invariante alle traslazioni se, nota l'uscita y ad un certo ingresso x, la sua uscita ad un ingresso costituito dal segnale x traslato nel tempo coincida con la versione traslata della y

$$y(n) = T\{x(n)\}$$

$$x(n - n_0) \Rightarrow y(n - n_0) \tag{A.3}$$

Rappresentando un segnale discreto x(n) come una serie di impulsi $\delta(n)$

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\, \delta(n - k) \tag{A.4}$$

dove

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{A.5}$$

per sistemi lineari tempo-invarianti, la relazione

$$y(n) = T\{ x(n) \} = T\left\{ \sum_{k=-\infty}^{\infty} x(k)\, \delta(n-k) \right\} \tag{A.6}$$

può essere espressa, per la linearità, come

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)\, T\{\delta(n-k)\} \tag{A.7}$$

Indicando con

$$h(n) = T\{\delta(n)\} \tag{A.8}$$

la risposta impulsiva del sistema, per l'invarianza alla traslazione si ottiene

$$h(n-k) = T\{\delta(n-k)\} \tag{A.9}$$

da cui

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)\, h(n-k) = x(n) \otimes h(n) \tag{A.10}$$

Si può quindi concludere che un sistema discreto lineare tempo invariante è completamente caratterizzato dalla sua risposta impulsiva o, equivalentemente, dalla funzione di trasferimento ottenuta come trasformata z della stessa

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{n=-\infty}^{\infty} h(n)\, z^{-n} \tag{A.11}$$

Lo svantaggio fondamentale della sommatoria di convoluzione introdotta per l'analisi dei sistemi nel dominio del tempo risiede nel fatto che gli estremi della sommatoria risultano pari a $\pm\infty$. Per evitare tale inconveniente, il legame ingresso-uscita può essere approssimato tramite equazioni alle

differenze. Una equazione lineare alle differenze a coefficienti costanti di ordine n è un legame tra tra gli ultimi M campioni dell'ingresso e gli ultimi N dell'uscita del tipo

$$\sum_{k=0}^{N} a_k\, y(n-k) = \sum_{k=0}^{M} b_k\, x(n-k) \qquad (A.12)$$

A seconda della struttura dell'equazione alle differenze, si possono distinguere tre differenti tipi di sistemi. Se N = 0 (sistema Moving Average: MA), l'equazione alle differenze si riduce a

$$y(n) = \sum_{k=0}^{M} \frac{b_k}{a_0}\, x(n-k) \qquad (A.13)$$

Tale equazione non è più ricorsiva, in quanto dipende esclusivamente dall'ingresso. Confrontando tale equazione con quella di definizione della risposta impulsiva di un sistema, si nota come i coefficienti b_k/a_0 coincidano con i campioni della risposta impulsiva h(k). Posto x(n) = δ (n) si ha che la risposta impulsiva è data da

$$h(n) = \sum_{k=0}^{N} \frac{b_k}{a_0}\, \delta(n-k) \qquad (A.14)$$

ed è di durata finita: il sistema è definito FIR (Finite Impulse Response). Se M = 0 (sistema Auto Regressive: AR), l'equazione alle differenze si semplifica come

$$y(n) = -\sum_{k=1}^{N} \frac{a_k}{a_0}\, y(n-k) + \frac{b_0}{a_0}\, x(n) \qquad (A.15)$$

A causa della retrazione dell'uscita, la risposta impulsiva del filtro è di durata finita (sistema IIR: Infinite Impulse Response). Il caso più generale è quello in cui sia M che N non sono nulli. Il sistema è quindi governato dall'equazione alle differenze

$$y(n) = \sum_{k=0}^{M} \frac{b_k}{a_0}\, x(n-k) - \sum_{k=1}^{N} \frac{a_k}{a_0}\, y(n-k) \qquad (A.16)$$

ed il sistema (di tipo IIR) è detto ARMA. Esprimendo l'equazione alle differenze

$$\sum_{k=0}^{N} a_k\, y(n-k) = \sum_{k=0}^{M} b_k\, x(n-k) \tag{A.17}$$

tramite la sua trasformata Z, si ottiene

$$Y(z) \sum_{k=0}^{N} a_k\, z^{k} = X(z) \sum_{k=0}^{M} b_k\, z^{-k} \tag{A.18}$$

con funzione di trasferimento

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k\, z^{-k}}{\sum_{k=0}^{N} a_k\, z^{-k}} \tag{A.19}$$

data da un rapporto di polinomi in z con $a_0 = 1$. Se $b_0 \neq 0$, è possibile eliminare le potenze negative di z fattorizzando i termini $b_0 z^{-M}$ e $a_0 z^{-N}$ ed ottenendo

$$H(z) = G\, z^{-M+N} \frac{\prod_{k=1}^{M} (z - z_k)}{\prod_{k=1}^{N} (z - p_k)} \tag{A.20}$$

dove $G = b_0/a_0 = b_0$. Per sistemi FIR il denominatore non è presente e la funzione di trasferimento, è composta da soli zeri. Per sistemi AR la funzione di trasferimento è composta da soli poli.

Essendo la funzione di trasferimento dipendente dal contributo dei singoli poli e zeri, è possibile prevedere le caratteristiche della risposta impulsiva di un sistema a tempo discreto (e, quindi, il suo comportamento nel dominio del tempo) in funzione della loro localizzazione. Mentre il contributo del numeratore è poco intuitivo (ma sarà analizzato nel seguito), l'influenza della localizzazione dei poli è, invece, molto evidente.

Imponendo il vincolo che l'uscita del sistema discreto sia una serie di valori reali, anche l'equazione alle differenze deve risultare a coefficienti reali. Nel caso di funzione di trasferimento con un solo polo, ciò comporta che il polo stesso si trovi sull'asse reale. Per quanto riguarda la risposta impulsiva si ha

$$H(z) = \frac{1}{1 - r z^{-1}} \quad \overset{z}{\leftrightarrow} \quad h(n) = r^n u(n) \tag{A.21}$$

Nel caso in cui il polo si trovi sul semiasse positivo ($r > 0$), la risposta impulsiva risulta decadere esponenzialmente, essere costante o crescere esponenzialmente a secondo che il suo modulo risulti minore, uguale o maggiore di 1. Il valore di r determina la rapidità del decadimento. Nel caso in cui il polo si trovi sul semiasse negativo ($r < 0$), l'inviluppo della risposta impulsiva ha lo stesso andamento precedentemente descritto in funzione del modulo, solo che i campioni risultano a segni alternati. Il comportamento non cambia sostanzialmente nel caso di poli reali coincidenti

$$H(z) = \frac{1}{\left(1 - r z^{-1}\right)^2} \quad \overset{z}{\leftrightarrow} \quad h(n) = n\, r^n u(n) \tag{A.22}$$

Per quanto riguarda poli complessi coniugati

$$H(z) = \frac{r \sin\omega_0\, z^{-1}}{\left(1 - r e^{j\omega_0} z^{-1}\right)\left(1 - r e^{-j\omega_0} z^{-1}\right)} \quad \overset{z}{\leftrightarrow} \quad h(n) = r^n \sin(\omega_0 n)\, u(n) \tag{A.23}$$

la risposta impulsiva è oscillante con inviluppo decrescente, costante o crescente, secondo che i poli si trovino all'interno, sulla o all'esterno della circonferenza di raggio unitario.

Se un sistema è stabile, l'uscita prodotta a seguito di un ingresso limitato deve risultare limitata. Essendo l'uscita funzione della risposta impulsiva, dall'analisi precedente segue che un sistema è stabile se ha poli esclusivamente all'interno della circonferenza unitaria.

Per quanto riguarda la risposta in frequenza del sistema discreto LTI, definita come trasformata della sua risposta impulsiva, si nota come essa coincida con la H(z) calcolata sulla circonferenza di raggio unitario

$$H(\omega) = \sum_{n=-\infty}^{\infty} h(n)\, e^{-j\omega n} = H(z)\,|_{z=e^{j\omega}}$$

$$H(\omega) = b_0\, e^{j\omega(N-M)} \frac{\prod_{k=1}^{M} \left(e^{j\omega} - z_k\right)}{\prod_{k=1}^{N} \left(e^{j\omega} - p_k\right)} \qquad (A.24)$$

Interpretando i termini nelle produttorie, $e^{j\omega}$ rappresenta il vettore uscente dall'origine con vertice sulla circonferenza unitaria nel punto di pulsazione ω, mentre z_k e p_k rappresentano i vettori che congiungono l'origine con gli zeri ed i poli (fig. A.1). I termini tra parentesi tonde, quindi, sono i vettori congiungenti i poli e gli zeri con la circonferenza, che in forma polare sono esprimibili come

$$\begin{cases} e^{j\omega} - z_k = V_k(\omega)\, e^{j\Theta_k(\omega)} \\ e^{j\omega} - p_k = U_k(\omega)\, e^{j\Phi_k(\omega)} \end{cases} \qquad (A.25)$$

con

$$\begin{cases} V_k(\omega) = |\, e^{j\omega} - z_k \,|\,; \quad \Theta_k(\omega) = \angle\left(e^{j\omega} - z_k\right) \\ U_k(\omega) = |\, e^{j\omega} - p_k \,|\,; \quad \Phi_k(\omega) = \angle\left(e^{j\omega} - p_k\right) \end{cases} \qquad (A.26)$$

Noti i moduli V_k e U_k dei vettori congiungenti i poli e gli zeri con la circonferenza unitaria e le loro rotazioni Θ_k e Φ_k, il modulo della funzione di trasferimento si ottiene come prodotto dei moduli

$$|\, H(\omega) \,| = |\, b_0 \,| \frac{\prod_{k=1}^{M} V_k(\omega)}{\prod_{k=1}^{N} U_k(\omega)} \qquad (A.27)$$

Analizzando il contributo dei singoli poli e zeri, si vede che il modulo della funzione di trasferimento si riduce per valori di ω corrispondenti a punti della circonferenza unitaria prossimi ad uno zero, in quanto si riduce il relativo termine V_k. Per quanto riguarda i poli, invece, la riduzione del temine U_k si

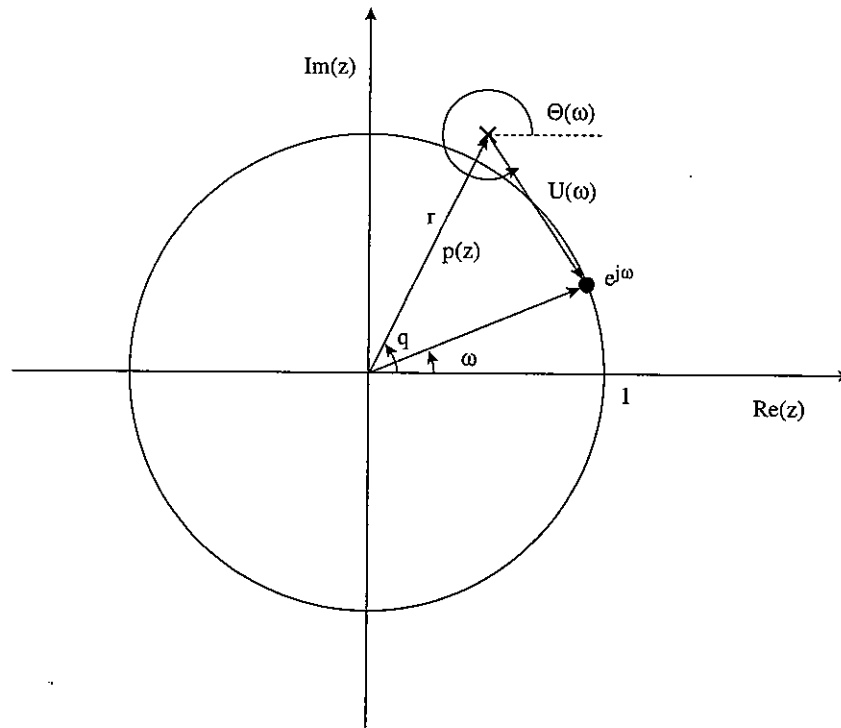



Fig. A.1 - Interpretazione geometrica del contributo di un polo al modulo ed alla fase della funzione di trasferimento.

traduce in un'esaltazione della risposta in frequenza, dato che tale termine si trova al denominatore della H(ω). È importante sottolineare come, a parità di modulo di due poli complessi coniugati, l'ampiezza del picco di risonanza dipenda dalla loro frequenza (fig. A.2). Viceversa, a parità di frequenza e interpretando il contributo di un polo come un filtraggio passa banda nell'intorno della sua frequenza centrale, la banda del filtro si riduca all'aumentare del modulo del polo. Tale miglioramento dal punto di vista della selettività in frequenza si paga, a causa della riduzione dello smorzamento, dal punto di vista della selettività temporale.

Per quanto riguarda la fase, essa dipende dalla differenza tra la somma delle fasi di ciascuno zero e quella dei poli

$$\angle H(\omega) = \angle b_0 + \omega \cdot (M-N) + \sum_{k=1}^{M} \Theta_k(\omega) - \sum_{k=1}^{N} \Phi_k(\omega) \qquad (A.28)$$

Nel caso di sistemi discreti LTI posti tra loro in cascata, si può verificare che la funzione di trasferimento complessiva è data dal prodotto delle funzioni di trasferimento dei singoli sistemi. Infatti

$$Y_2(z) = X_2(z) \cdot H_2(z) = Y_1(z) \cdot H_2(z) = X_1(z) \cdot H_1(z) \cdot H_2(z) \quad (A.29)$$

Ipotizzando di voler realizzare il sistema inverso ad uno dato. Ciò vuol dire che si vuole realizzare un sistema che messo in cascata ad un secondo produce una funzione di trasferimento complessiva unitaria. A tal fine è immediato verificare che è sufficiente realizzare un sistema che abbia come poli gli zeri del sistema da invertire e viceversa. Affinché anche il sistema inverso sia stabile, è necessario che i suoi poli siano all'interno della circonferenza unitaria, il che si traduce con il vincolo che il sistema originario, oltre ad avere i propri poli all'interno della circonferenza unitaria per la sua stabilità, vi abbia anche gli zeri. Sistemi con sia poli che zeri all'interno della circonferenza unitaria sono detti a fase minima.

A.2 FUNZIONE DI AUTOCORRELAZIONE DI UN PROCESSO AUTOREGRESSIVO

L'equazione caratteristica di un processo AR, soddisfatta sia dal processo u(n)

$$u(n) = v(n) - \sum_{k=1}^{q} w_k \, u(n-k) \quad (A.30)$$

che dall'autocorrelazione R(n)

$$\sum_{k=0}^{q} w_k \, R_{uu}(n-k) = 0 \, ; \quad \begin{cases} n = 1, 2, \ldots, q \\ w_0 = 1 \end{cases} \quad (A.31)$$

è pari a

$$1 + \sum_{k=1}^{q} w_k \, z^{-k} = 0 \quad (A.32)$$

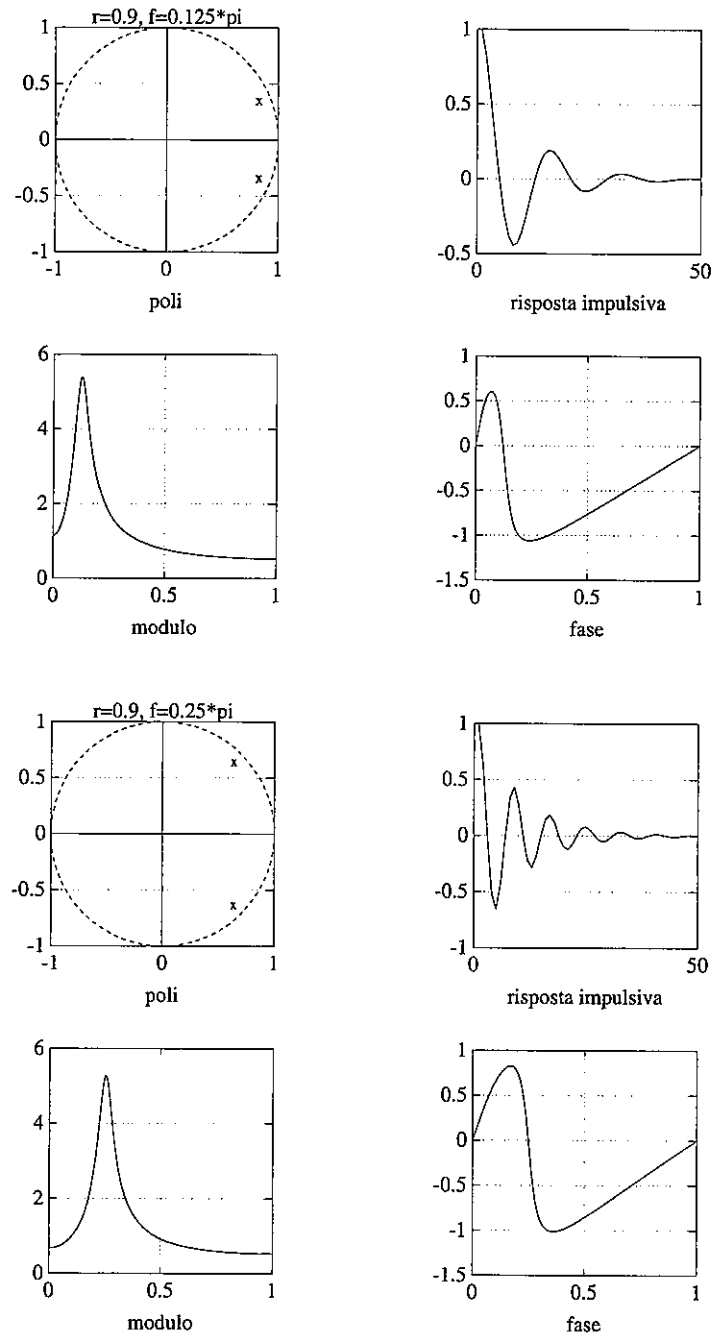



Fig. A.2a - Legame tra posizione dei poli, risposta impulsiva e risposta in frequenza.

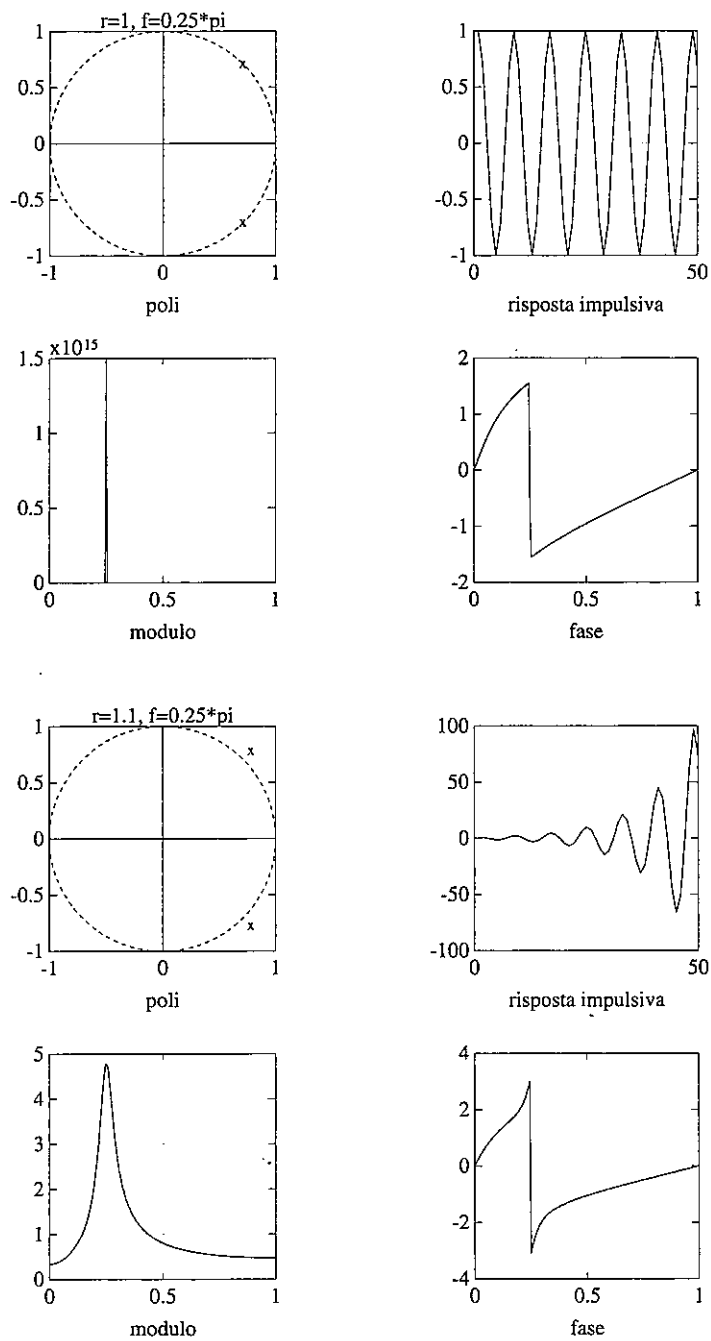



Fig. A.2b - Legame tra posizione dei poli, risposta impulsiva e risposta in frequenza.

Per la stabilità, le radici p dell'equazione caratteristica debbo risiedere all'interno del circolo di raggio unitario. Si vuole mostrare come, variando i coefficienti del sistema AR, si riesca a variare le caratteristiche della funzione di autocorrelazione del processo generato.

Per un processo del secondo ordine, la funzione caratteristica è pari a

$$1 + w_1 z^{-1} + w_2 z^{-2} = 0 \tag{A.33}$$

con poli

$$p_1, p_2 = \frac{-w_1 \pm \sqrt{w_1^2 - 4w_2}}{2} \quad \begin{cases} |p_1| < 1 \\ |p_2| < 1 \end{cases} \rightarrow \begin{cases} w_2 + w_1 \geq -1 \\ w_2 - w_1 \geq 1 \\ -1 \leq w_2 \leq 1 \end{cases} \tag{A.34}$$

Imponendone la stabilità, si ricava la zona di esistenza dei coefficienti (fig. A.3)

$$\begin{cases} |p_1| < 1 \\ |p_2| < 1 \end{cases} \rightarrow \begin{cases} w_2 + w_1 \geq -1 \\ w_2 - w_1 \geq 1 \\ -1 \leq w_2 \leq 1 \end{cases} \tag{A.35}$$

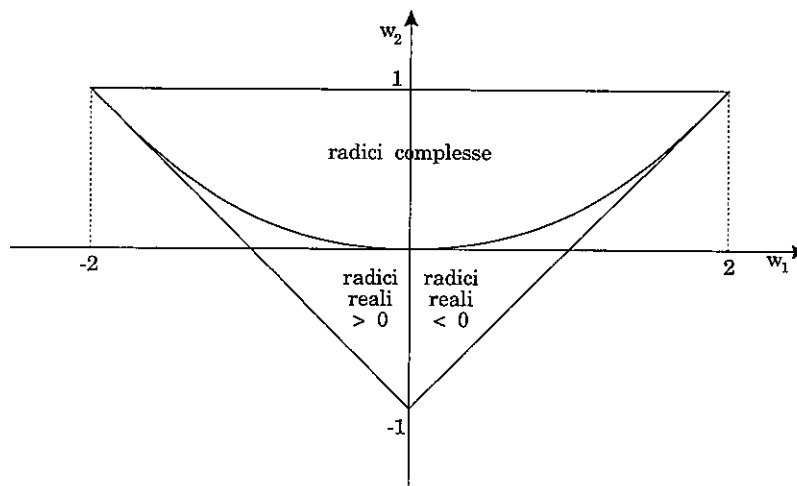



Fig. A.3 - Regione di esistenza per i coefficienti del processo AR.

All'interno di tale zona, cambiando i coefficienti del processo, la funzione di autocorrelazione assume le seguenti caratteristiche

$$\begin{cases} w_1^2 - 4\,w_2 > 0 \;\rightarrow\; \text{radici reali} \begin{cases} w_1 < 0 \;\rightarrow\; \begin{array}{l}\text{radice dominante positiva;}\\ \text{decadimento esponenziale}\end{array} \\ w_1 > 0 \;\rightarrow\; \begin{array}{l}\text{radice dominante negativa;}\\ \text{decadimento esponenziale a segni alterni}\end{array} \end{cases} \\ w_1^2 - 4\,w_2 < 0 \;\rightarrow\; \text{radici complesse coniugate} \;\rightarrow\; \text{oscillazioni} \end{cases} \tag{A.36}$$

L'andamento della funzione di autocorrelazione è riportata in figura A.4 per i tre casi.

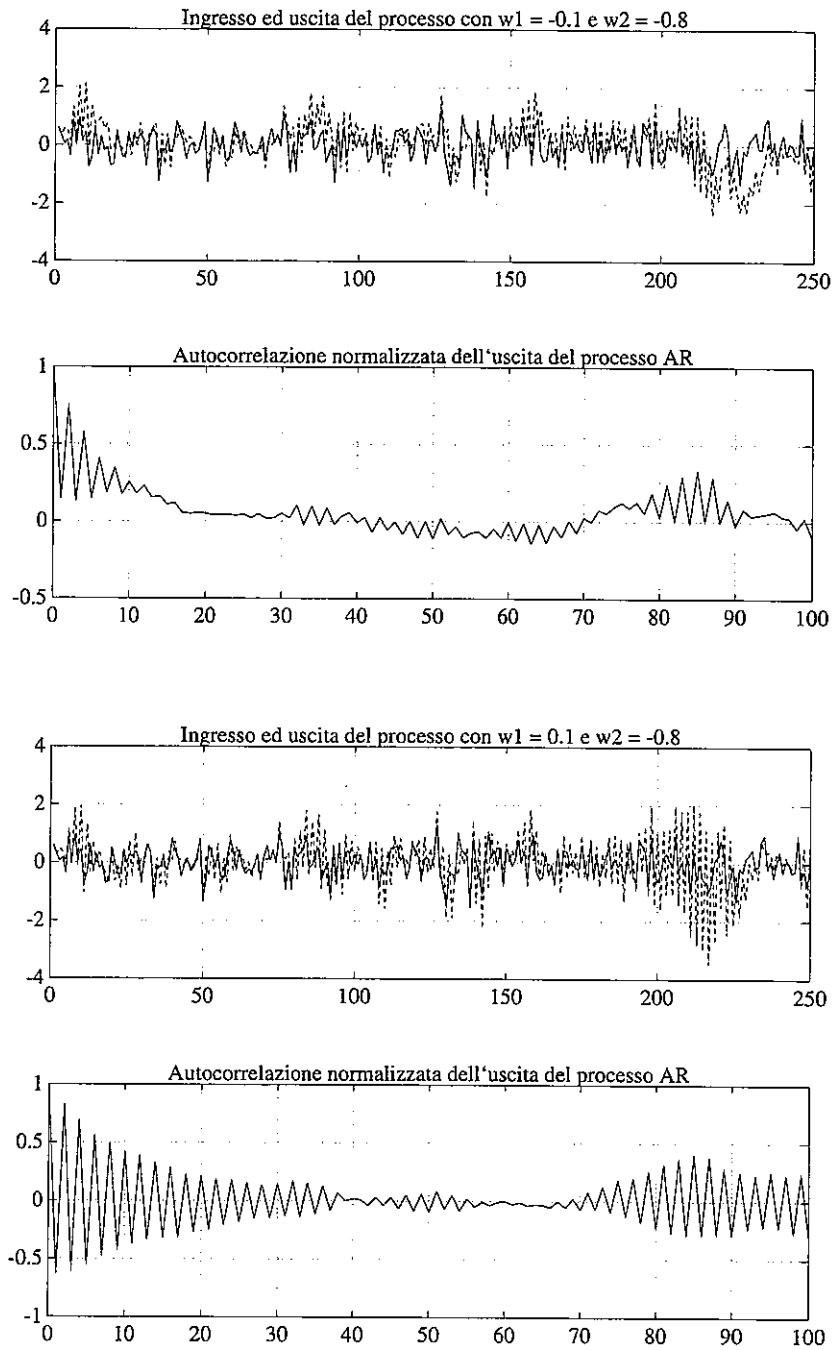



Fig. A.4a - Funzioni di autocorrelazione di processi autoregressivi.

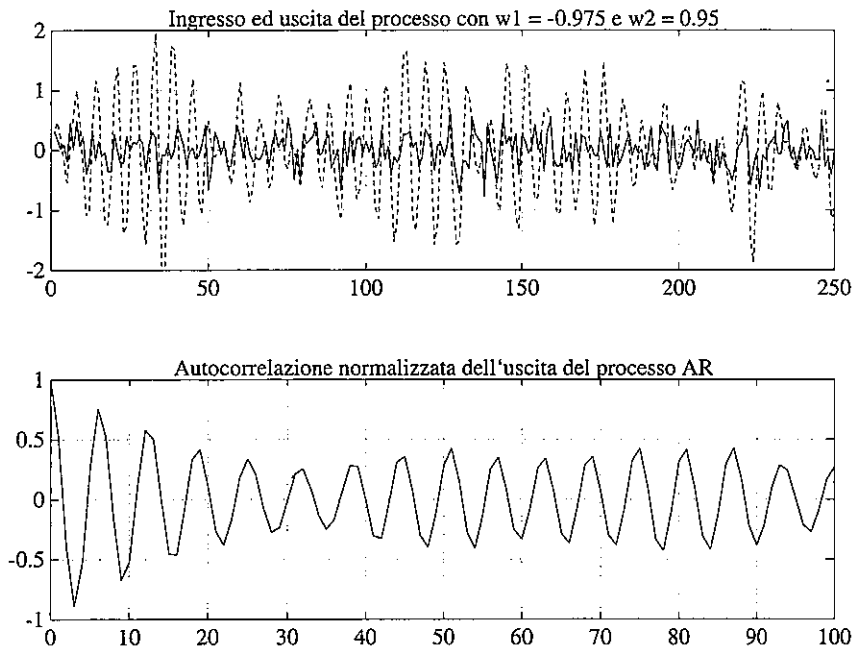



Fig. A.4b - Funzioni di autocorrelazione di processi autoregressivi.

Appendice B

METODO DEI MINIMI QUADRATI RICORSIVO

Il vantaggio principale dell'LMS è la sua semplicità. Per contro, il mantenere fisso il passo di aggiornamento è svantaggioso in quanto esso risulta essere solitamente troppo piccolo nelle fasi iniziali, ritardando la convergenza, ed eccessivo nelle fasi finali, deteriorando le prestazioni. Inoltre, dipendendo l'LMS solamente dall'errore istantaneo, non ha memoria dei precedenti passi di elaborazione. È possibile introdurre un differente algoritmo, derivato dall'LS, in cui l'aggiornamento dei coefficienti del predittore avviene indipendentemente e ricorsivamente per ciascuno di essi in funzione dell'andamento della stima (Recursive Least-Square: RLS). Ciò introduce memoria nell'algoritmo, come mostrato nel seguito, ma questo non rappresenta uno svantaggio nel caso di fenomeni stazionari. Nel caso di processi variabili, invece, è necessario attenuare la memoria sui campioni più remoti man mano che si procede nella stima. Ciò si ottiene con l'adozione di un nuovo criterio di ottimizzazione che introduce un fattore di pesatura $w < 1$ nella definizione dell'errore quadratico

$$\hat{\varepsilon}(n) = \sum_{n=0}^{t} w^{t-n} e^{(n)2} \tag{B.1}$$

È poi necessario introdurre la funzione di errore a priori, definita come

$$\eta^{(n)} = x(n) - \alpha^{(n-1)T} \mathbf{x}(n-1) \tag{B.2}$$

nella quale si stima il campione corrente in funzione dei pesi calcolati nel passo precedente, che risulta solitamente differente dalla funzione di errore a posteriori

$$e^{(n)} = x(n) - \alpha^{(n)T} x(n-1) \tag{B.3}$$

nella quale si utilizza il vettore dei coefficienti aggiornato. Ripetendo i passi già presentati nel capitolo 5, la minimizzazione dell'errore in funzione dei coefficienti del predittore porta al passo n-esimo all'equazione normale

$$\Gamma^{(n)} \alpha^{(n)} = \theta^{(n)} \tag{B.4}$$

che risulta formalmente identica a quella utilizzata nella derivazione delle equazioni di Wiener-Hopf, ma nelle quale è implicitamente presente il fattore di pesatura w. La matrice di autocorrelazione dell'ingresso ed il vettore di cross-correlazione tra l'ingresso e l'uscita desiderata risultano, infatti, pari a

$$\Gamma^{(n)} = \sum_{i=1}^{n} w^{n-i}\, \mathbf{x}(i-1)\, \mathbf{x}^{T}(i-1)$$

$$\theta^{(n)} = \sum_{i=1}^{n} w^{n-i}\, x(i)\, \mathbf{x}^{T}(i-1) \tag{B.5}$$

La soluzione di tale sistema è ancora dato da

$$\alpha^{(n)} = \Gamma^{(n)^{-1}} \theta^{(n)} \tag{B.6}$$

anche se, in questo caso, ne ricerca una sua soluzione ricorsiva. Per quanto riguarda la matrice di autocorrelazione, è possibile darne una definizione ricorsiva isolandone dalla definizione il termine per i = n

$$\Gamma^{(n)} = w \left[\sum_{i=0}^{n-1} w^{n-1-i}\, \mathbf{x}(i-1)\, \mathbf{x}^{T}(i-1) \right] + \mathbf{x}(n-1)\, \mathbf{x}^{T}(n-1)$$

$$\Gamma^{(n)} = w\, \Gamma^{(n-1)} + \mathbf{x}(n-1)\, \mathbf{x}^{T}(n-1) \tag{B.7}$$

Analogamente si ottiene

$$\theta^{(n)} = w\,\theta^{(n-1)} + x(n)\,\mathbf{x}^{T}(n-1) \tag{B.8}$$

L'inversione della matrice di autocorrelazione può essere evitata, in quanto risulta [Hay86]

$$\Gamma^{(n)^{-1}} = \frac{\Gamma^{(n-1)^{-1}}}{w} - \frac{\dfrac{\Gamma^{(n-1)^{1}}\,\mathbf{x}(n-1)\,\mathbf{x}^{T}(n-1)\,\Gamma^{(n-1)^{-1}}}{w^{2}}}{1 + \dfrac{\mathbf{x}^{T}(n-1)\,\Gamma^{(n-1)^{-1}}\,\mathbf{x}(n-1)}{w}} \tag{B.9}$$

Per semplificare la notazione si introduce la matrice di autocorrelazione inversa

$$\mathbf{P}^{(n)} = \Gamma^{(n)^{1}} \tag{B.10}$$

ed il vettore dei guadagni

$$\mathbf{k}^{(n)} = \frac{\dfrac{\mathbf{P}^{(n-1)}\,\mathbf{x}(n-1)}{w}}{1 + \dfrac{\mathbf{x}^{T}(n-1)\,\mathbf{P}^{(n-1)}\,\mathbf{x}(n-1)}{w}} \tag{B.11}$$

che ha le stesse dimensioni dell'ordine del predittore. In tal modo, per l'equazione dell'inversa della matrice di autocorrelazione si ottiene la relazione ricorsiva

$$\mathbf{P}^{(n)} = \frac{\mathbf{P}^{(n-1)} - \mathbf{k}^{(n)}\,\mathbf{x}^{T}(n-1)\,\mathbf{P}^{(n-1)}}{w} \tag{B.12}$$

Riorganizzando la definizione del vettore dei guadagni si ottiene, invece

$$\mathbf{k}^{(n)} = \frac{\left[\mathbf{P}^{(n-1)} - \mathbf{k}^{(n)}\,\mathbf{x}^{T}(n-1)\,\mathbf{P}^{(n-1)}\right]\mathbf{x}(n-1)}{w} = \mathbf{P}^{(n)}\,\mathbf{x}(n-1) \tag{B.13}$$

con la quale si ottengono le seguenti relazioni ricorsive per i coefficienti del predittore

$$\alpha^{(n)} = \mathbf{P}^{(n)}\,\theta^{(n)} = w\,\mathbf{P}^{(n)}\,\theta^{(n-1)} + x(n)\,\mathbf{P}^{(n)}\,\mathbf{x}(n-1)$$

$$= \mathbf{P}^{(n-1)}\,\theta^{(n-1)} - \mathbf{k}^{(n)}\,\mathbf{x}^T(n)\,\mathbf{P}^{(n-1)}\,\theta^{(n-1)} + x(n)\,\mathbf{P}^{(n)}\,\mathbf{x}(n-1)$$

$$= \alpha^{(n-1)} + x(n)\,\mathbf{P}^{(n)}\,\mathbf{x}(n-1) - \mathbf{k}^{(n)}\,\mathbf{x}^T(n-1)\,\alpha^{(n-1)}$$

$$= \alpha^{(n-1)} + \mathbf{k}^{(n)}\left[\, x(n) - \mathbf{x}^T(n-1)\,\alpha^{(n-1)} \,\right] \qquad \text{(B.14)}$$

Sostituendo l'espressione dell'errore

$$\eta^{(n)} = x(n) - \mathbf{x}^T(n-1)\,\alpha^{(n-1)} \qquad \text{(B.15)}$$

si ottiene, infine

$$\alpha^{(n)} = \alpha^{(n-1)} + \mathbf{k}^{(n)}\,\eta^{(n)} \qquad \text{(B.16)}$$

Da tale equazione si nota come il vettore dei guadagni k(n) ha le stesse funzioni del parametro μ nell'LMS. A differenza di quest'ultimo, però, esso risulta variabile e, dipendendo dalla matrice di correlazione del segnale, ha memoria del processo di stima.

In tal modo sono state ottenute equazioni ricorsive per il calcolo di k, μ, α e P. Al fine di ridurre la complessità computazionale, però, è opportuno definire la grandezza

$$\mathbf{y}^{(n)} = \frac{\mathbf{P}^{(n-1)}\,\mathbf{x}(n-1)}{w} \qquad \text{(B.17)}$$

con la quale i passi per la predizione sono ridotti ai seguenti:

– si calcola la grandezza

$$\mathbf{y}^{(n)} = \frac{\mathbf{P}^{(n-1)}\,\mathbf{x}(n-1)}{w} \qquad \text{(B.18)}$$

– da questa si calcola il vettore dei guadagni

$$\mathbf{k}^{(n)} = \frac{\mathbf{y}^{(n)}}{1 + \mathbf{x}^T(n-1)\,\mathbf{y}^{(n)}} \qquad \text{(B.19)}$$

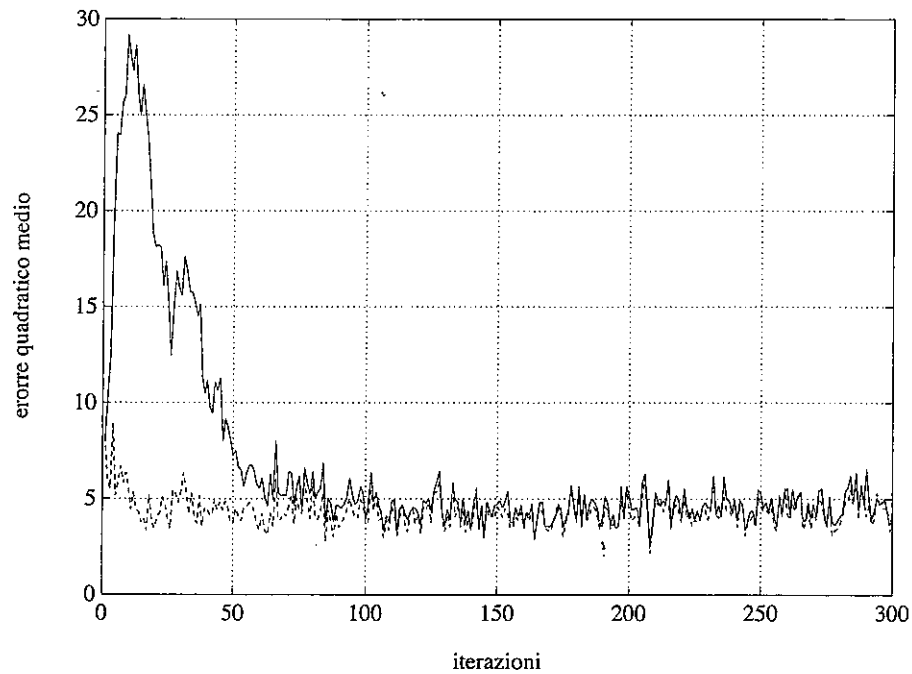



Fig. B.1 - Curve di apprendimento per LMS e RLS.

– si calcola l'errore

$$\eta^{(n)} = x(n) - \mathbf{x}^T(n-1)\,\alpha^{(n-1)} \tag{B.20}$$

– si aggiornano i parametri del predittore

$$\alpha^{(n)} = \alpha^{(n-1)} + \eta^{(n)}\,\mathbf{k}^{(n)} \tag{B.21}$$

– si calcola, infine, il valore aggiornato della matrice di autocorrelazione

$$\mathbf{P}^{(n)} = \frac{\mathbf{P}^{(n-1)}}{w} - \mathbf{k}^{(n)}\,\mathbf{y}^{(n)} \tag{B.22}$$

Per quanto riguarda le condizioni iniziali, è necessario scegliere la matrice P(0) in modo tale che risulti non singolare. La scelta più immediata è quella di scegliere una costante δ positiva "piccola" (es.: 0.1) e fissare

$$\mathbf{P}^{(0)} = \frac{\mathbf{I}}{\delta} \tag{B.23}$$

Per i coefficienti del predittore il loro valore iniziale è tipicamente nullo, mentre il coefficiente di smorzamento w deve risultare eventualmente minore, ma comunque molto prossimo ad 1 (es.: $1 - 10^{-3}$).

Dal punto di vista delle prestazioni, la variabilità del vettore dei guadagni fa si che sia i tempi di convergenza che le fluttuazioni nell'intorno del vettore ottimo siano ridotti rispetto all'LMS (fig. B.1). In compenso, però, la complessità computazionale è aumentata, anche considerando versioni più sofisticate dell'algoritmo [Hay86].

Appendice C

ALGORITMI A BLOCCHI PER IL FILTRAGGIO ADATTATIVO

Gli algoritmi per la soluzione dell'equazione normale della predizione a blocchi possono essere ricondotti, fondamentalmente, alle seguenti due classi (fig. C.1)

- metodo della covarianza;
- metodo dell'autocorrelazione;

La loro differenza risiede nel differente modo con il quale vengono stimati gli elementi della matrice di autocorrelazione Φ. La stima della funzione di auto correlazione viene calcolata come sommatoria dei prodotti ottenuti utilizzando due finestre di campioni di pari lunghezza progressivamente traslate. Nel secondo caso si utilizza un'unica finestra di campioni. Le traslazioni richieste nel calcolo dei prodotti avvengono estendendo tale finestra base con dei campioni nulli. Nel primo caso, invece, la finestra base di campioni viene estesa utilizzando effettivi campioni del segnale. Tale differenza porta poi ad una differente struttura della matrice Φ e quindi a differenti algoritmi per la soluzione del sistema lineare di equazioni associato.

Il metodo dell'autocorrelazione è quello attualmente più utilizzato sia per l'esistenza di algoritmi più robusti nei riguardi di errori di quantizzazione (adottando strutture a traliccio), sia per l'esistenza di più efficienti algoritmi di calcolo (ricorsione di Schür).

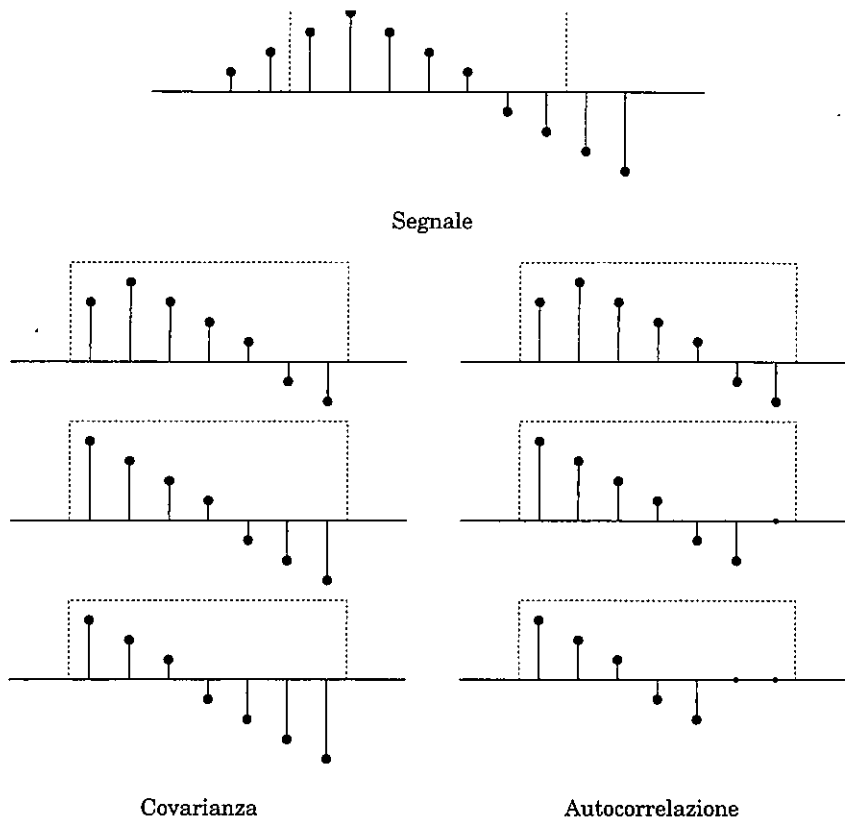



Fig. C.1 -Differenti finestre utilizzate nel metodo della covarianza e dell'autocorrelazione.

C.1 METODO DELLA COVARIANZA

Nel metodo della covarianza gli elementi della matrice Φ vengono stimati utilizzando sequenze di "n" campioni del segnale, progressivamente sfasate. Ciascun elemento della matrice viene cioè calcolato come

$$\Phi(i,j) = \frac{1}{n}\sum_{k=0}^{n-1} x(k-i)\,x(k-j); \quad \begin{array}{l} 1 \le i \le p \\ 1 \le j \le p \end{array}$$

$$= \frac{1}{n}\sum_{k=-i}^{n-i-1} x(k)\,x(k+i-j) \qquad \text{(C.1)}$$

Considerare blocchi di "n" campioni del segnale vuol dire moltiplicare il segnale stesso per una finestra di pari lunghezza. La finestra utilizzata nel seguito è quella rettangolare, tralasciando considerazioni sull'opportunità di utilizzare finestre differenti.

Per il calcolo degli elementi $\Phi(i, k)$ da utilizzare nella predizione, è necessario considerare il prodotto di sequenze traslate nel tempo al massimo di "p" campioni. Se la finestra di interesse è quella relativa alla sequenza x(n), è necessario considerare fino a "p" campioni esterni a tale intervallo. Con tale approssimazione, la matrice Φ dell'equazione normale deterministica viene indicata come matrice di covarianza, da cui il nome dell'algoritmo. Si mette in evidenza come tale nome non indica che i termini della matrice rappresentino la covarianza del segnale, ma sono sempre termini di autocorrelazione. La soluzione dell'equazione si ottiene poi come

$$\alpha = \Phi^{-1} \psi \tag{C.2}$$

Es.: si consideri il vettore

$$x(n) = [2 \ -1 \ 0 \ -1 \ 2] \tag{C.3}$$

ottenuto considerando la funzione f = cos(t) + cos(2t) e prelevando 5 campioni nel suo periodo. Considerando un predittore del secondo ordine, l'equazione normale è

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \Phi(1,1) & \Phi(1,2) \\ \Phi(2,1) & \Phi(2,2) \end{bmatrix}^{-1} \begin{bmatrix} \Phi(1,0) \\ \Phi(2,0) \end{bmatrix} \tag{C.4}$$

Per il calcolo degli elementi della matrice di covarianza è necessario estendere l'intervallo di osservazione, includendo anche due campioni del precedente periodo. Nel nostro caso traslazioni verso destra o verso sinistra si equivalgono.

$$\begin{aligned} x(n-1) &= [\,-1 \ \ 0 \ -1 \ \ 2 \ -1\,] \\ x(n-2) &= [\ \ 0 \ -1 \ \ 2 \ -1 \ \ 0\,] \end{aligned} \tag{C.5}$$

È possibile, quindi, calcolare gli elementi della matrice di covarianza

$$\Phi(1,0)=\frac{1}{5}\left([-1\ 0\ -1\ 2\ -1]\,[2\ -1\ 0\ -1\ 2]^T\right)=-\frac{6}{5}$$
$$\Phi(2,0)=\frac{1}{5}\left([0\ -1\ 2\ -1\ 0]\,[2\ -1\ 0\ 1\ 2]^T\right)=\frac{2}{5}$$
$$\Phi(1,1)=\frac{1}{5}\left([-1\ 0\ -1\ 2\ -1]\,[-1\ 0\ -1\ 2\ -1]^T\right)=\frac{7}{5}$$
$$\Phi(2,2)=\frac{1}{5}\left([0\ -1\ 2\ -1\ 0]\,[0\ -1\ 2\ -1\ 0]^T\right)=\frac{6}{5}$$
$$\Phi(2,1)=\frac{1}{5}\left([0\ -1\ 2\ -1\ 0]\,[-1\ 0\ -1\ 2\ 1]^T\right)=-\frac{4}{5} \qquad (C.6)$$

e calcolare il vettore dei coefficienti ottimi come

$$\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=\begin{bmatrix}\frac{7}{5} & -\frac{4}{5}\\ -\frac{4}{5} & \frac{6}{5}\end{bmatrix}^{-1}\begin{bmatrix}-\frac{6}{5}\\ \frac{2}{5}\end{bmatrix}=\begin{bmatrix}\frac{15}{13} & \frac{10}{13}\\ \frac{10}{13} & \frac{35}{26}\end{bmatrix}\begin{bmatrix}-\frac{6}{5}\\ \frac{2}{5}\end{bmatrix}=\begin{bmatrix}-\frac{14}{13}\\ -\frac{5}{13}\end{bmatrix} \qquad (C.7)$$

Il calcolo della Φ può avvenire anche evitando di utilizzare campioni esterni al blocco corrente. A tal fine è necessario considerare sequenze di n-p campioni opportunamente sfasate ritagliate all'interno del blocco. Il calcolo della Φ può poi avvenire costruendo la matrice delle osservazioni

$$\mathbf{H}^T=\begin{bmatrix} x(p) & x(p{+}1) & \dots & x(n)\\ x(p-1) & x(p-1) & \dots & x(n-1)\\ \dots & \dots & \dots & \dots\\ x(1) & x(2) & \dots & x(n-p+1)\end{bmatrix} \qquad (C.8)$$

dalla quale si ricavano la matrice di auto-correlazione ed il vettore di cross-correlazione

$$\Phi=\frac{\mathbf{H}^T\mathbf{H}}{\mathbf{n}-\mathbf{p}}$$

$$\psi=\frac{\mathbf{H}^T\mathbf{x(n{+}1)}}{n-p} \qquad (C.9)$$

La matrice di covarianza risulta essere definita positiva e, per costruzione, simmetrica. La soluzione dell'equazione normale può essere quindi trovata evitando l'inversione della stessa, ma scomponendola tramite l'algoritmo di Cholesky nel prodotto

$$\Phi = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^{T} \qquad (C.10)$$

dove **V** è una matrice triangolare inferiore con la diagonale principale unitaria, mentre **D** è una matrice diagonale

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ V_{21} & 1 & 0 & \dots & 0 \\ V_{31} & V_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ V_{p1} & V_{p2} & V_{p3} & \dots & 1 \end{bmatrix}; \quad \mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & d_p \end{bmatrix} \qquad (C.11)$$

Per eseguire la fattorizzazione della matrice di covarianza si procede come segue: dato

$$\Phi(i, j) = \sum_{k=1}^{j} V_{ik}\, d_k\, V_{jk}; \quad 1 \le j \le i - 1 \qquad (C.12)$$

per gli elementi sulla diagonale segue

$$d_i = \Phi(i, i) - \sum_{k=1}^{i-1} V_{ik}^2\, d_k; \quad i \ge 2 \qquad (C.13)$$

a partire dalla condizione iniziale

$$d_1 = \varphi(1, 1) \qquad (C.14)$$

mentre gli elementi della matrice **V** si ricavano riscrivendo l'espressione del generico elemento della matrice di covarianza isolando dalla sommatori il termine per k=j

$$V_{ij}\, d_j = \Phi(i, j) - \sum_{k=1}^{j-1} V_{ik}\, d_k\, V_{jk}; \quad 1 \le j \le i - 1$$

$$V_{ij} = \frac{\Phi(i, j) - \sum_{k=1}^{j-1} V_{ik}\, d_k\, V_{jk}}{d_j} \qquad (C.15)$$

In tal modo è possibile calcolare ricorsivamente gli elementi V_{ij} e d_j. Una volta fattorizzata la matrice di covarianza, anche la soluzione dell'equazione normale può essere ottenuta ricorsivamente. Infatti, definendo il vettore **y** come

$$\mathbf{y} = \mathbf{D}\,\mathbf{V}^T \alpha \tag{C.16}$$

risulta

$$\Phi\,\alpha = \mathbf{V}\,\mathbf{D}\,\mathbf{V}^T \alpha = \mathbf{V}\,\mathbf{y}$$

$$\mathbf{V}\,\mathbf{y} = \psi \tag{C.17}$$

Essendo **V** una matrice triangolare, gli elementi della **y** possono essere calcolati ricorsivamente come

$$y_1 = \psi_1$$

$$y_i = \psi_i - \sum_{j=1}^{i-1} V_{ij}\, y_j; \quad p \ge i \ge 2 \tag{C.18}$$

Una volta calcolati i componenti del vettore **y** è possibile ricavare i coefficienti ottimi del predittore in quanto, risultando

$$\mathbf{V}^T \alpha = \mathbf{D}^{-1}\,\mathbf{y} \tag{C.19}$$

Si ottengono le relazioni ricorsive

$$\alpha_i = \frac{y_i}{d_i} - \sum_{j=i+1}^{p} V_{ji}\,\alpha_j; \quad 1 \le i \le p-1 \tag{C.20}$$

con condizioni iniziali

$$\alpha_p = \frac{y_p}{d_p} \tag{C.21}$$

Es. si consideri lo stesso vettore di campioni precedente. Applicando la fattorizzazione di Cholesky risulta

$$d_1 = \Phi(1, 1) = \frac{7}{5}$$

$$V(2, 1) = \frac{\Phi(2, 1)}{d_1} = -\frac{4}{7}$$

$$d_2 = \Phi(2, 2) - V(2, 1)^2 d_1 = \frac{26}{35} \tag{C.22}$$

Fattorizzata la matrice di covarianza come

$$\mathbf{\Phi = V\,D\,V^T}$$

$$\Phi = \begin{bmatrix} 1 & 0 \\ \frac{4}{7} & 1 \end{bmatrix} \begin{bmatrix} \frac{7}{5} & 0 \\ 0 & \frac{26}{35} \end{bmatrix} \begin{bmatrix} 1 & -\frac{4}{7} \\ 0 & 1 \end{bmatrix} \tag{C.23}$$

si calcolano dapprima le variabili ausiliarie y

$$\mathbf{V\,y = \psi}$$

$$\begin{bmatrix} 1 & 0 \\ -\frac{4}{7} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} -\frac{6}{5} \\ \frac{2}{5} \end{bmatrix} \tag{C.24}$$

tramite la relazione

$$y_i = \psi_i - \sum_{j=1}^{i-1} V_{ij}\, y_{j;} \quad p \geq i \geq 2$$

$$y_1 = \Phi(1, 0) = -\frac{6}{5}$$

$$y_2 = \Phi(2, 0) - V(2, 1)\, y_1 = -\frac{2}{7} \tag{C.25}$$

ed infine i coefficienti del predittore dalla relazione

$$\alpha_i = \frac{y_i}{d_i} - \sum_{j=i+1}^{p} V_{ji}\, \alpha_j; \quad 1 \le i \le p-1$$

$$\alpha_2 = \frac{y_2}{d_2} = -\frac{5}{13}$$

$$\alpha_1 = \frac{y_1}{d_1} - V(2,1)\,\alpha_2 = -\frac{14}{13} \tag{C.26}$$

C.2 METODO DELL'AUTOCORRELAZIONE

Il metodo dell'autocorrelazione stima gli elementi della matrice F utilizzando esclusivamente campioni all'interno di una finestra di lunghezza "n". Ciò equivale ad assumere implicitamente che il segnale sia nullo al di fuori di tale intervallo e quindi la traslazione di sequenze di campioni si ottengono con il suo completamento tramite campioni nulli. In tal modo il calcolo dell'autocorrelazione è eseguito su di un numero di campioni non nulli via via inferiore man mano che si calcolano valori di autocorrelazione di indice crescente. Gli elementi della matrice di autocorrelazione andrebbero quindi calcolati come

$$R(i) = \frac{1}{n-i} \sum_{k=1}^{n-i} x(k)\, x(k+i); \quad 1 \le i \le p \tag{C.27}$$

In tal modo, però, al diminuire del denominatore, verrebbe esaltato lo scostamento della stima dal valore effettivo dell'autocorrelazione, scostamento dovuto al ridursi del numero di campioni utili. Si preferisce, quindi, non eseguire il calcolo degli elementi della matrice della autocorrelazione (non polarizzata) secondo tale definizione, ma calcolare gli elementi della matrice della autocorrelazione (polarizzata) come

$$R(i) = \frac{1}{n} \sum_{k=1}^{n-i} x(k)\, x(k+i); \quad 1 \le i \le p \tag{C.28}$$

Ciò permette di attenuare il contributo degli elementi esterni in quanto, al ridursi delle dimensioni delle sequenze, il valore del denominatore rimane costante. L'equazione normale deterministica in forma matriciale diventa

$$\begin{bmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \dots \\ R(p) \end{bmatrix}$$

$$\mathbf{R}\,\alpha = \mathbf{r} \qquad \text{(C.29)}$$

con soluzione

$$\alpha = \mathbf{R}^{-1}\,\mathbf{r} \qquad \text{(C.30)}$$

Es.: ripetiamo il calcolo dei coefficienti del predittore del secondo ordine

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} R(0) & R(1) \\ R(1) & R(0) \end{bmatrix}^{1} \begin{bmatrix} R(1) \\ R(2) \end{bmatrix} \qquad \text{(C.31)}$$

per lo stesso segnale precedentemente considerato. Gli elementi della matrice di autocorrelazione sono dati da

$$\begin{aligned} &x(n) = [2\ -1\ 0\ -1\ 2] && R(0) = \frac{1}{5}\left([2\ -1\ 0\ -1\ 2]\,[2\ -1\ 0\ -1\ 2]^T \right) = 2 \\ &x(n+1) = [-1\ 0\ -1\ 2\ 0] && R(1) = \frac{1}{5}\left([2\ -1\ 0\ -1\ 2]\,[-1\ 0\ 1\ 2\ 0]^T \right) = \frac{4}{5} \\ &x(n+2) = [0\ -1\ 2\ 0\ 0] && R(2) = \frac{1}{5}\left([2\ -1\ 0\ -1\ 2]\,[0\ -1\ 2\ 0\ 0]^T \right) = \frac{1}{5} \end{aligned} \qquad \text{(C.32)}$$

I coefficienti del predittore sono, quindi, calcolati come

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 2 & -\frac{4}{5} \\ -\frac{4}{5} & 2 \end{bmatrix}^{-1} \begin{bmatrix} -\frac{4}{5} \\ \frac{1}{5} \end{bmatrix} = \begin{bmatrix} \frac{25}{42} & \frac{5}{21} \\ \frac{5}{21} & \frac{25}{42} \end{bmatrix} \begin{bmatrix} -\frac{4}{5} \\ \frac{1}{5} \end{bmatrix} = \begin{bmatrix} -\frac{9}{21} \\ -\frac{3}{42} \end{bmatrix} \qquad \text{(C.33)}$$

Anche in questo caso è possibile utilizzare la matrice delle osservazioni, avente dimensioni [n + m - 1, m], che assume la forma

$$\mathbf{H}^T = \begin{bmatrix} x(1) & x(2) & \dots & x(p) & x(p+1) & \dots & x(n) & 0 & \dots & 0 \\ 0 & x(1) & \dots & x(p-1) & x(p-1) & \dots & x(n-1) & x(n) & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x(1) & x(2) & \dots & x(n-p+1) & x(n-p) & \dots & x(n) \end{bmatrix} \tag{C.34}$$

Dalla H si ricavano la matrice di autocorrelazione ed il vettore di cross-correlazione

$$\mathbf{R} = \frac{\mathbf{H}^T \mathbf{H}}{\mathbf{n}}$$

$$\mathbf{r} = \frac{\mathbf{H}^T \mathbf{x(n+1)}}{n} \tag{C.35}$$

La soluzione diretta dell'equazione di Yule-Walker tramite l'inversione della matrice di auto correlazione può essere anche in questo caso evitata, dato che essa risulta essere una matrice di Toeplitz, cioè simmetrica con tutti gli elementi di una diagonale identici. Ciò permette di ricavare ricorsivamente i coefficienti di un predittore di ordine p in funzione dei coefficienti di predittori di ordine inferiore tramite una relazione ricorsiva del tipo

$$\alpha_k^{(m)} = \alpha_k^{(m-1)} + \delta_k^{(m-1)} \tag{C.36}$$

dove $\alpha_k^{(m)}$ indica il coefficiente k-esimo del predittore di ordine "m" e quindi al passo m-esimo dell'algoritmo. Le condizioni iniziali sono quelle di un predittore del primo ordine, che possono essere ricavate direttamente dall'equazione normale

$$R(0)\,\alpha_1^{(1)} = R(1)$$

$$\alpha_1^{(1)} = \frac{R(1)}{R(0)} \tag{C.37}$$

Dall'espressione del minimo dell'MSE, per un predittore del primo ordine si ricava

$$e^{(1)} = R(0) - \alpha_1^{(1)} R(1) = R(0)\left\{1 - \left[\alpha_1^{(1)}\right]^2\right\} \qquad (C.38)$$

Considerando poi il problema di calcolare i coefficienti di un predittore di ordine m in funzione dei coefficienti di un predittore di ordine (m-1) è necessario poter risolvere il sistema

$$\alpha^{(m)} = \begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \dots \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \alpha^{(m-1)} \\ 0 \end{bmatrix} + \begin{bmatrix} \delta^{(m-1)} \\ \hline k^{(m)} \end{bmatrix} \qquad (C.39)$$

con $\delta^{(m-1)}$ e $k^{(m)}$ da determinare. In tal modo si ottiene l'equazione aumentata del predittore

$$\begin{bmatrix} \mathbf{R}^{(m-1)} & \mathbf{r}_b^{(m-1)} \\ \mathbf{r}_b^{(m-1)T} & R(0) \end{bmatrix} \left\{ \begin{bmatrix} \alpha^{(m-1)} \\ 0 \end{bmatrix} + \begin{bmatrix} \delta^{(m-1)} \\ \hline k^{(m)} \end{bmatrix} \right\} = \begin{bmatrix} \mathbf{r}^{(m-1)} \\ R(m) \end{bmatrix} \qquad (C.40)$$

dove $\mathbf{r}_b^{(m-1)}$ rappresenta il vettore di cross-correlazione al passo m-1 $\mathbf{r}^{(m-1)}$ con gli elementi in ordine invertito. Da questa relazione matriciale derivano le due relazioni

$$\mathbf{R}^{(m-1)}\,\alpha^{(m-1)} + \mathbf{R}^{(m-1)}\,\delta^{(m-1)} + k^{(m)}\,\mathbf{r}_b^{(m-1)} = \mathbf{r}^{(m-1)}$$

$$\mathbf{r}_b^{(m-1)T}\,\alpha^{(m-1)} + \mathbf{r}_b^{(m-1)T}\,\delta^{(m-1)} + k^{(m)}\,R(0) = R(m) \qquad (C.41)$$

Per l'equazione del predittore di ordine (m - 1)

$$\mathbf{R}^{(m-1)}\,\alpha^{(m-1)} = \mathbf{r}^{(m-1)} \qquad (C.42)$$

dalla prima equazione si ottiene

$$\delta^{(m-1)} = -k^{(m)}\,\mathbf{R}^{(m-1)^{-1}}\,\mathbf{r}_b^{(m-1)}$$

$$\delta^{(m-1)} = -k^{(m)}\,\alpha_b^{(m-1)} \qquad (C.43)$$

dove $\alpha_b^{(m-1)}$ rappresenta il vettore dei coefficienti del predittore di ordine m-1 $\alpha^{(m-1)}$ con gli elementi in ordine invertito. Sostituendo l'espressione di $\delta^{(m-1)}$ nell'equazione iniziale è possibile ricavare ricorsivamente i coefficienti del predittore come

$$\alpha^{(m)} = \alpha^{(m-1)} - k^{(m)}\,\alpha_b^{(m-1)}$$

$$\alpha_j^{(m)} = \alpha_j^{(m-1)} - k^{(m)}\,\alpha_{m-j}^{(m-1)},\quad 1 \le j \le m-1 \tag{C.44}$$

In tale espressione la costante $k^{(m)}$ risulta ancora incognita. Per poterla ricavare si sostituisce l'espressione di $\delta^{(m-1)}$ nella seconda componente dell'equazione normale aumentata, ottenendo

$$\mathbf{r}_b^{(m-1)^T}\,\alpha^{(m-1)} - k^{(m)}\,\mathbf{r}_b^{(m-1)^T}\,\alpha_b^{(m-1)} + k^{(m)}\,R(0) = R(m)$$

$$k^{(m)} = \frac{R(m) - \mathbf{r}_b^{(m-1)^T}\,\alpha^{(m-1)}}{R(0) - \mathbf{r}_b^{(m-1)^T}\,\alpha_b^{(m-1)}} = \frac{R(m) - \mathbf{r}_b^{(m-1)^T}\,\alpha^{(m-1)}}{R(0) - \mathbf{r}_b^{(m-1)^T}\,\mathbf{R}^{(m-1)^{-1}}\,\mathbf{r}_b^{(m-1)}} = \frac{R(m) - \mathbf{r}_b^{(m-1)^T}\,\alpha^{(m-1)}}{R(0) - \alpha^{(m-1)^T}\,\mathbf{r}^{(m-1)}} \tag{C.45}$$

Riconoscendo nel denominatore l'espressione dell'errore minimo al passo m-1, si ha

$$k^{(m)} = \frac{R(m) - \mathbf{r}_b^{(m-1)^T}\,\alpha^{(m-1)}}{e^{(m-1)}} = \frac{R(m) - \sum_{j=1}^{m-1} \alpha_j^{(m-1)}\,R(m-j)}{e^{(m-1)}} \tag{C.46}$$

Anche l'errore può essere calcolato ricorsivamente, dato che

$$e^{(m)} = R(0) - \alpha^{(m-1)^T}\,R_{m-1} = R(0) - \sum_{k=1}^{m} \alpha_k^{(m)}\,R(k) \tag{C.47}$$

Sostituendo l'espressione ricorsiva degli α_k si ottiene

$$e^{(m)} = R(0) - \sum_{k=1}^{m-1} \alpha_k^{(m-1)}\,R(k) - \alpha_m^{(m)}\left[R(m) - \sum_{k=1}^{m-1} \alpha_{m-k}^{(m-1)}\,R(k) \right] \tag{C.48}$$

Confrontando l'espressione in parentesi quadre con il numeratore dell'espressione che fornisce $k^{(m)}$, si ottiene

$$e^{(m)} = e^{(m-1)} - \alpha_m^{(m)} k^{(m)} e^{(m-1)} \tag{C.49}$$

ed essendo $k^{(m)} = \alpha_m^{(m)}$, si ha, infine

$$e^{(m)} = \left[1 - k^{(m)^2}\right] e^{(m-1)} \tag{C.50}$$

Riepilogando, è possibile calcolare ricorsivamente tramite l'algoritmo di Levinson-Durbin i coefficienti di un predittore di ordine p in funzione dei coefficienti di predittori di ordine inferiore. Ciò si ottiene tramite i seguenti passi:

– si determina il valore dell'errore dalla relazione

$$e^{(m)} = R(0) - \sum_{k=1}^{m} \alpha_k^{(m)} R(k)$$

$$\text{con valore iniziale } e^{(0)} = R(0) \tag{C.51}$$

– si ricava il valore del coefficiente $k^{(m)}$ dalla relazione

$$k^{(m)} = \frac{R(m) - \sum_{j=1}^{m-1} \alpha_j^{(m-1)} R(m-j)}{e^{(m-1)}}$$

$$\text{con valore iniziale } k^{(1)} = \frac{R(1)}{e^{(0)}} \tag{C.52}$$

– si determinano i coefficienti del predittore

$$\alpha_m^{(m)} = k^{(m)}.$$

$$\alpha_j^{(m)} = \alpha_j^{(m-1)} - k^{(m)} \alpha_{m-j}^{(m-1)}, \quad 1 \le j \le m-1 \tag{C.53}$$

– si determina il valore aggiornato dell'errore

$$e^{(m)} = \left[1 - k^{(m)^2}\right] e^{(m-1)} \tag{C.54}$$

Tale procedura si ripete per un numero di passi pari all'ordine del predittore.

Es.: ripetiamo il calcolo dei coefficienti del predittore per lo stesso segnale precedentemente considerato. L'errore al passo iniziale è

$$e^{(0)} = R(0) = 2 \tag{C.55}$$

Il ciclo per m = 1 porta al calcolo delle grandezze di un predittore del primo ordine

$$k^{(1)} = \frac{R(1)}{e^{(0)}} = -\frac{2}{5}$$

$$\alpha_1^{(1)} = k^{(1)} = -\frac{2}{5}$$

$$e^{(1)} = \left(1 - k^{(1)^2}\right) e^{(0)} = \frac{42}{25} \tag{C.56}$$

Per m = 2 si ottengono le grandezze del predittore del secondo ordine

$$k^{(2)} = \frac{R(2) - \alpha_1^{(1)} R(1)}{e^{(1)}} = -\frac{3}{42}$$

$$\alpha_2^{(2)} = k^{(2)} = -\frac{3}{42}$$

$$\alpha_1^{(2)} = \alpha_1^{(1)} - k^{(2)} \alpha_1^{(1)} = -\frac{9}{21} \tag{C.57}$$

C.3 METODI UTILIZZANTI STRUTTURE A TRALICCIO

Sia il metodo della covarianza che il metodo dell'autocorrelazione precedentemente esposti sono concettualmente organizzati su due fasi: una stima preliminare della matrice di autocorrelazione del segnale ed il successivo calcolo dei coefficienti del predittore. Il metodo che si vuole presentare in questo paragrafo (lattice) può essere visto come un raffinamento del metodo dell'autocorrelazione, nel quale tali due fasi sono fuse in un'unica procedura ricorsiva.

Per descrivere l'algoritmo è necessario riprendere l'equazione del predittore

$$\hat{x}(n) = \sum_{k=1}^{p} \alpha_k \, x(n-k) \qquad (C.58)$$

e della sua funzione di trasferimento

$$P(z) = \sum_{k=1}^{p} \alpha_k \, z^{-k} \qquad (C.59)$$

Definita la funzione d'errore del predittore di ordine "p" come

$$e^{(p)}(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} \alpha_k^{(p)} \, x(n-k) \qquad (C.60)$$

è possibile definire la funzione di trasferimento del sistema d'errore. Si considera, cioè, il sistema che, dato come ingresso il segnale, produce in uscita la funzione d'errore. Tale funzione di trasferimento è pari a

$$A^{(p)}(z) = 1 - \sum_{k=1}^{p} \alpha_k \, z^{-k} \qquad (C.61)$$

In tal modo è possibile esprimere l'errore di predizione in z come

$$E^{(p)}(z) = A^{(p)}(z) \, X(z) \qquad (C.62)$$

Ora ci si può porre il problema di esprimere la A(z) di un predittore di ordine "m" in funzione della A(z) di un predittore di ordine "m-1". Per tale problema si può utilizzare la trattazione svolta per l'algoritmo di Levinson-Durbin che fornisce un'espressione ricorsiva per i coefficienti del predittore α_k. Sostituendo tale relazione nella definizione di A(z) si ottiene

$$A^{(m)}(z) = 1 - \sum_{j=1}^{m} \alpha_j^{(m)} \, z^{-j} = 1 - \sum_{j=1}^{m} \left[\alpha_j^{(m-1)} + k^{(m)} \, \alpha_{m-j}^{(m-1)} \right] z^{-j}$$

$$= 1 - \sum_{j=1}^{m-1} \left[\alpha_j^{(m-1)} - k^{(m)} \alpha_{m-j}^{(m-1)} \right] z^j - \alpha_m^{(m)} z^m$$

$$= \left[1 - \sum_{j=1}^{m-1} \alpha_j^{(m-1)} z^{-j} \right] - \left[\alpha_m^{(m)} z^{-m} - \sum_{j=1}^{m-1} k^{(m)} \alpha_{m-j}^{(m-1)} z^{-j} \right]$$

$$A^{(m)}(z) = A^{(m-1)}(z) - k^{(m)} z^m A^{(m-1)}\left(z^{-1}\right) \tag{C.63}$$

Sostituendo nell'espressione dell'errore si ottiene

$$E^{(m)}(z) = A^{(m-1)}(z)\, X(z) - k^{(m)} z^{-m} A^{(m-1)}(z^{-1})\, X(z) \tag{C.64}$$

Il primo termine del secondo membro rappresenta la trasformata z della funzione d'errore di un predittore di ordine "m - 1". Tale predittore stima il campione x(n) utilizzando i campioni [x(n - m) ... x(n - 1)] e verrà nel seguito indicato come predittore in avanti. Per il secondo termine si può dare un'interpretazione simile se si definisce

$$B^{(m)}(z) = z^{-m} A^{(m)}(z^{-1})\, X(z) \tag{C.65}$$

che ha un'antitrasformata del tipo

$$b^{(m)}(n) = x(n - m) - \sum_{k=1}^{m} \alpha_k^{(m)} x(n + k - m) \tag{C.66}$$

Questo è interpretabile come l'errore di un predittore che tenta di stimare x(n - m) in funzione degli "m - 1" campioni [x(n m 1) ... x(n)]. Tale predittore verrà nel seguito indicato come predittore all'indietro. Sostituendo la B(z) nella E(z) si ottiene

$$E^{(m)}(z) = E^{(m-1)}(z) - k^{(m)} B^{(m-1)}(z^{-1}) \tag{C.67}$$

che, antitrasformata fornisce

$$e^{(m)}(n) = e^{(m-1)}(n) - k^{(m)} b^{(m-1)}(n - 1) \tag{C.68}$$

Tale espressione rappresenta una relazione ricorsiva della funzione d'errore della predizione in avanti. Sostituendo l'espressione ricorsiva della A(z) nella definizione della B(z) si ottiene

$$B^{(m)}(z) = z^{-m} A^{(m-1)}(z^{-1}) X(z) - k^{(m)} A^{(m-1)}(z) X(z)$$

$$= z^{-1} B^{(m-1)}(z) X(z) - k^{(m)} E^{(m-1)}(z) \qquad (C.69)$$

che, antitrasformata, fornisce la relazione ricorsiva della funzione d'errore della predizione all'indietro

$$b^{(m)}(n) = b^{(m-1)}(n-1) - k^{(m)} e^{(m-1)}(n) \qquad (C.70)$$

Se si traccia il diagramma di flusso di tale algoritmo si ottiene una struttura a traliccio (lattice), da cui il nome dell'algoritmo (fig. C.2). È quindi possibile il calcolo ricorsivo delle due funzioni d'errore. Le condizioni iniziali sono quelle per un predittore di ordine zero, cioè senza alcuna predizione e quindi

$$e^{(0)}(n) = b^{(0)}(n) = x(n) \qquad (C.71)$$

In forma matriciale le relazioni ricorsive in z sono esprimibili come

$$\begin{bmatrix} A^{(m)}(z) \\ B^{(m)}(z) \end{bmatrix} = \begin{bmatrix} 1 & k^{(m)} \\ k^{(m)} z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} A^{(m-1)}(z) \\ B^{(m-1)}(z) \end{bmatrix} \qquad (C.72)$$

Per il calcolo dei coefficienti $k^{(m)}$ si impone la simultanea minimizzazione degli errori quadratici di predizione in avanti e all'indietro

$$\begin{cases} E\left\{\left[e^{(m-1)}(n)\right]^2\right\} = \sum_{i=0}^{n-1}\left[e^{(m-1)}(i)\right]^2 \\ E\left\{\left[b^{(m-1)}(n)\right]^2\right\} = \sum_{i=0}^{n-1}\left[b^{(m-1)}(i)\right]^2 \end{cases} \qquad (C.73)$$

La minimizzazione, ottenuta annullando le derivate rispetto ai k

$$\begin{cases} \dfrac{\partial E\left\{\left[e^{(m)}(n)\right]^2\right\}}{\partial k} = 0 \\ \dfrac{\partial E\left\{\left[b^{(m)}(n)\right]^2\right\}}{\partial k} = 0 \end{cases}$$

$$\begin{cases} k^{(m)} E\left\{\left[b^{(m-1)}(n)\right]^2\right\} = E\left\{e^{(m-1)}(n)\; b^{(m-1)}(n-1)\right\} \\ k^{(m)} E\left\{\left[e^{(m-1)}(n)\right]^2\right\} = E\left\{e^{(m-1)}(n)\; b^{(m-1)}(n-1)\right\} \end{cases} \tag{C.74}$$

porta alla seguente espressione ricorsiva per il calcolo dei coefficienti

$$k^{(m)} = \frac{\sum_{i=0}^{n-1}\left[e^{(m-1)}(i)\; b^{(m-1)}(i-1)\right]}{\sqrt{\sum_{i=0}^{n-1}\left[e^{(m-1)}(i)\right]^2 \sum_{i=0}^{n-1}\left[b^{(m-1)}(i)\right]^2}} \tag{C.75}$$

Dato che tale espressione ha la forma di una cross-correlazione tra le funzioni di errore di predizione in avanti e all'indietro, i coefficienti $k^{(m)}$ vengono detti PARCOR (PARtial CORrelation). Noti i coefficienti $k^{(m)}$ è possibile determinare i coefficienti del predittore come già visto nel Levinson-Durbin (dal quale il presente algoritmo deriva)

$$\alpha_m^{(m)} = k^{(m)}.$$

$$\alpha_j^{(m)} = \alpha_j^{(m-1)} - k^{(m)}\, \alpha_{m-j}^{(m-1)} \quad 1 \le j \le m-1 \tag{C.76}$$

Per chiarire ulteriormente la struttura degli errori di predizione è opportuno analizzare con un esempio la loro struttura. Per un predittore di ordine zero risulta

$$\begin{cases} e_0(n) = x(n) \\ b_0(n) = x(n) \end{cases} \tag{C.77}$$

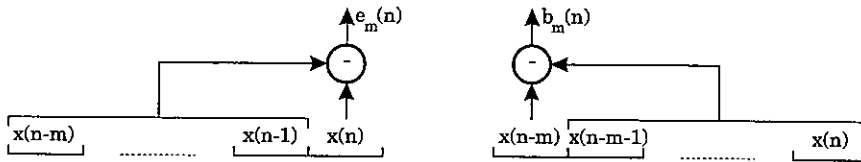
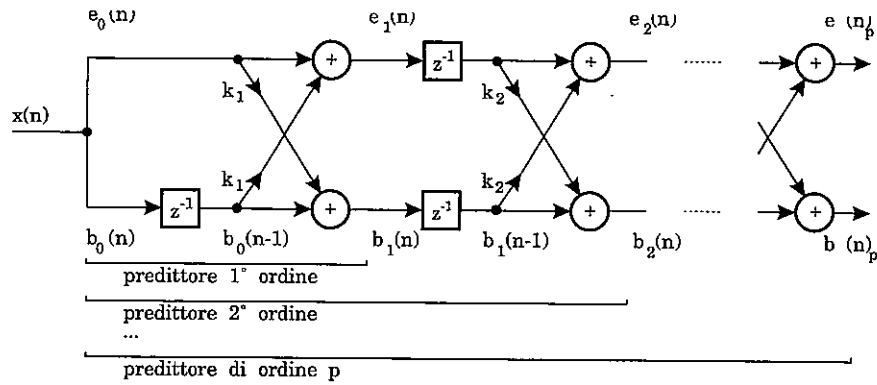



Fig. C.2 - Predittore di ordine p implementato tramite una struttura a traliccio.

Per un predittore di ordine uno la predizione in avanti utilizza solamente il campione che precede il corrente, mentre la predizione all'indietro utilizza il campione corrente per predire il precedente. Infatti risulta

$$\begin{cases} e^{(1)}(n) = e^{(0)}(n) + k^{(1)}\, b^{(0)}(n-1) = x(n) + \alpha_1^{(1)}\, x(n-1) \\ b^{(1)}(n) = k^{(1)}\, e^{(0)}(n) + b^{(0)}(n-1) = \alpha_1^{(1)}\, x(n) + x(n-1) \end{cases} \quad (C.78)$$

Analogamente, per un predittore di ordine due risulta

$$\begin{cases} e^{(2)}(n) = e^{(1)}(n) + k^{(2)}\, b^{(1)}(n-1) = x(n) + \alpha_1^{(2)}\, x(n-1) + \alpha_2^{(2)}\, x(n-2) \\ b^{(2)}(n) = k^{(2)}\, e^{(1)}(n) + b^{(1)}(n-1) = \alpha_2^{(2)}\, x(n) + \alpha_1^{(2)}\, x(n-1) + x(n-2) \end{cases}$$

(C.79)

e così via.

Riassumendo, i passi per il metodo lattice sono i seguenti:

- si impostano le condizioni iniziali

$$e^{(0)}(n) = b^{(0)}(n) = x(n) \tag{C.80}$$

- si calcolano i coefficienti PARCOR

$$k^{(m)} = \frac{\sum_{m=0}^{n-1}\left[e^{(m-1)}(n) \quad b^{(m-1)}(n-1)\right]}{\sqrt{\sum_{m=0}^{n-1}\left[e^{(m-1)}(n)\right]^2 \sum_{m=0}^{n-1}\left[b^{(m-1)}(n)\right]^2}} \tag{C.81}$$

- si calcolano i coefficienti del predittore

$$\alpha_m^{(m)} = k^{(m)}$$

$$\alpha_j^{(m)} = \alpha_j^{(m-1)} - k^{(m)}\,\alpha_{m-j}^{(m-1)} \quad 1 \le j \le m-1 \tag{C.82}$$

- si aggiornano le funzioni di errore

$$e^{(m)}(n) = e^{(m-1)}(n) - k^{(m)}\,b^{(m-1)}(n-1)$$

$$b^{(m)}(n) = b^{(m-1)}(n-1) - k^{(m)}\,e^{(m-1)}(n) \tag{C.83}$$

con vettori che, inizialmente di dimensioni pari ad "n", assumono dimensioni crescenti fino a "n + p".

Es.: si ripete il calcolo del predittore del secondo ordine con il segnale precedentemente utilizzato. Le condizioni iniziali sono

$$x(n) = [2\ -1\ 0\ -1\ 2]$$

$$e^{(0)}(n) = [2\ -1\ 0\ -1\ 2]$$

$$b^{(0)}(n) = [2\ -1\ 0\ -1\ 2]$$

$$b^{(0)}(n-1) = [0\ 2\ -1\ 0\ 1\ 2] \tag{C.84}$$

Il coefficiente PARCOR del primo ordine è

$$k^{(1)} = \frac{\sum e^{(0)}(n)\, b^{(0)}(n-1)}{\sqrt{\sum e^{(0)}(n)^2\, b^{(0)}(n)^2}} = \frac{\sum [2\ -1\ 0\ -1\ 2]\,[0\ 2\ -1\ 0\ -1]}{\sqrt{\sum [2\ -1\ 0\ -1\ 2]^2 \sum [2\ -1\ 0\ -1\ 2]^2}}$$

$$= \frac{\sum [0\ -2\ 0\ 0\ -2]}{\sqrt{\sum [4\ 1\ 0\ 1\ 4] \sum [4\ 1\ 0\ 1\ 4]}} = -\frac{2}{5} \tag{C.85}$$

che porta alle nuove funzioni di errore

$$e^{(1)}(n) = e^{(0)}(n) - k^{(1)}\, b^{(0)}(n-1) = [2\ -1\ 0\ -1\ 2\ 0] + \frac{2}{5}\,[0\ 2\ -1\ 0\ -1\ 2]$$

$$= \left[2\ \frac{-1}{5}\ \frac{2}{5}\ -1\ \frac{8}{5}\ \frac{4}{5}\right]$$

$$b^{(1)}(n) = b^{(0)}(n-1) - k^{(1)}\, e^{(0)}(n) = [0\ 2\ -1\ 0\ -1] + \frac{2}{5}\,[2\ -1\ 0\ 1\ 2]$$

$$= \left[\frac{4}{5}\ \frac{8}{5}\ 1\ \frac{-2}{5}\ \frac{-1}{5}\ 2\right] \tag{C.86}$$

ed al coefficiente

$$\alpha_1^{(1)} = k^{(1)} = -\frac{2}{5} \tag{C.87}$$

Per il secondo ordine si ha

$$k^{(2)} = \frac{\sum e^{(1)}(n)\, b^{(1)}(n-1)}{\sqrt{\sum e^{(1)}(n)^2 \ \sum b^{(1)}(n)^2}} = \frac{\sum \left[2\ \frac{-1}{5}\ \frac{-2}{5}\ -1\ \frac{8}{5}\ \frac{4}{5}\right]\left[0\ \frac{4}{5}\ \frac{8}{5}\ -1\ \frac{-2}{5}\ \frac{-1}{5}\right]}{\sqrt{\sum \left[2\ \frac{-1}{5}\ \frac{-2}{5}\ -1\ \frac{8}{5}\ \frac{4}{5}\right]^2 \left[\frac{4}{5}\ \frac{8}{5}\ -1\ \frac{-2}{5}\ \frac{-1}{5}\ 2\right]^2}} = -\frac{3}{42} \tag{C.88}$$

Si possono quindi calcolare i coefficienti del predittore del secondo ordine

$$\alpha_2^{(2)} = k^{(2)} = -\frac{3}{42}$$

$$\alpha_1^{(2)} = \alpha_1^{(1)} - k^{(2)}\, \alpha_1^{(1)} = -\frac{9}{21} \tag{C.89}$$

Si noti come i valori dei coefficienti coincidono con quelli calcolati con il metodo dell'auto-correlazione

C.4 METODO RICORSIVO DI SCHÜR

Si osserva che l'algoritmo di Levinson-Durbin fa uso di una relazione ricorsiva per l'errore al passo m-esimo in funzione dell'errore e dei coefficienti PARCOR $k^{(m)}$ al passo (m-1) del tipo

$$e^{(m)} = [1 - (k^{(m)})^2]\, e^{(m-1)} \tag{C.90}$$

È possibile esprimere tale errore il funzione dell'autocorrelazione esplicitandola come

$$e^{(1)} = [1 - (k^{(1)})^2]\, R(0)$$

$$e^{(2)} = [1 - (k^{(1)})^2]\, [1 - (k^{(m)})^2]\, R(0) = [1 - (k^{(1)})^2]^2\, R(0)$$

$$\dots$$

$$e^{(m)} = [1 - (k^{(1)})^2]^m\, R(0) \tag{C.91}$$

Nell'algoritmo di Schür si esprimono tutte le relazioni al generico passo m-esimo in funzione degli elementi della matrice di autocorrelazione R(i) (i=0,1,...,m) e dei parametri $k^{(i)}$ (i=1,2,...,m). Per l'algoritmo di Schür si definiscono innanzitutto gli elementi

$$G_0 = \begin{bmatrix} 0 & R(1) & R(2) & \dots & R(p) \\ R(0) & R(1) & R(2) & \dots & R(p) \end{bmatrix}$$

$$G_1 = \begin{bmatrix} 0 & R(1) & R(2) & \dots & R(p) \\ 0 & R(0) & R(1) & \dots & R(p-1) \end{bmatrix}$$

$$k^{(1)} = \frac{G_1(1,2)}{G_1(2,2)}$$

$$V_1 = \begin{bmatrix} 1 & -k^{(1)} \\ k^{(1)} & 1 \end{bmatrix} \tag{C.92}$$

dove le dimensioni di G_0 sono pari a 2 * (p+1), con p ordine del predittore. Si esegue poi il prodotto V_1G_1 e si definisce G_2 come il risultato della precedente operazione con la seconda riga traslata di un posto verso destra. Si ripete il procedimento per p passi, ottenendo al generico passo m-esimo le seguenti relazioni ricorsive

$$k^{(m)} = \frac{G_m(1, m+1)}{G_m(2, m+1)}$$

$$V_m = \begin{bmatrix} 1 & -k^{(m)} \\ k^{(m)} & 1 \end{bmatrix} \quad \text{(C.93)}$$

Seguendo passo-passo lo svolgimento dell'algoritmo di Schür, si può facilmente verificare che al passo m-esimo l'elemento $G_m(2,m+1)$, ovvero il denominatore della relazione che definisce $K^{(m)}$, corrisponde al termine $e^{(m-1)}$ nell'algoritmo di Levinson-Durbin. Ad esempio, se implementiamo la stima dei parametri di un predittore di ordine 3 la matrice G_0 è definita come

$$G_0 = \begin{bmatrix} 0 & R(1) & R(2) & R(3) \\ R(0) & R(1) & R(2) & R(3) \end{bmatrix} \quad \text{(C.94)}$$

Si ottiene poi:

$$G_1(2,2) = R(0) = e^{(0)}$$

$$G_2(2,3) = R(0) - k^{(1)} R(1) = [1 - (k^{(1)})^2] R(0) = e^{(1)}$$

$$G_3(2,4) = R(0) - k^{(1)} R(1) - k^{(2)} [R(2) - k^{(1)} R(1)] = R(0) [1 - (k^{(1)})^2][1 - (k^{(2)})^2] = e^{(2)}$$

$$\text{(C.95)}$$

dove, al generico passo m-esimo, si è ricavato R(m) invertendo la relazione che definisce $k^{(m)}$ nell'algoritmo di Levinson Durbin.

Si nota anche che al generico passo m-esimo dell'algoritmo di Schür, la matrice G_m presenta le prime m colonne nulle, il che può essere sfruttato per ridurre sia la complessità di calcolo che l'occupazione di memoria.

Infatti, la versione dell'algoritmo di Schür proposta dallo standard GSM organizza le informazioni contenute nelle matrici G_m in due vettori (K e P) ed evita, ad ogni passo, di aggiornarne quegli elementi che non apportano

contributo ai passi successivi. In questo modo non si utilizzano relazioni matriciali; tutte le operazioni sono ridefinite sugli elementi di K e di P, vettori rispettivamente di dimensione (p-1) e (p+1). Al generico passo m, la determinazione di r(m) è sempre effettuata come rapporto tra il secondo elemento di P ed il primo; vengono aggiornati tutti gli elementi di P e di K tranne gli ultimi m.

Appendice D

RICHIAMI SU FILTRI NUMERICI

D.1 STRUTTURE PER MULTIRATE DSP

Per Multirate Digital Signal Processing si intende la realizzazione di sistemi discreti tali che il periodo T' dei segnali in uscita sia differente da quello T dei segnali in ingresso. Nel caso in cui la frequenza di campionamento del segnale in uscita sia maggiore di quella del segnale d'ingresso, è necessaria un'operazione di interpolazione per ricostruire il segnale x(t) mancante. Viceversa, nel caso in cui la frequenza del segnale in uscita sia minore di quella d'ingresso, è necessario decimare il segnale, eliminando campioni. È, però, necessario far precedere alla decimazione un'operazione di filtraggio che limiti la banda del segnale da ricampionare, al fine di evitare aliasing. Indicando con h(n) la risposta impulsiva dei filtri coinvolti nel cambiamento di frequenza (anti aliasing o di interpolazione), l'uscita è prodotta come

$$y(m) = x(t)\,|_{t=mT'} = \sum_{n=-\infty}^{\infty} x(n)\; h(t-nT)\,|_{t=mT'} = \sum_{n=-\infty}^{\infty} x(n)\; h(mT'-nT) \qquad \text{(D.1)}$$

In un sistema discreto è conveniente che il rapporto tra i periodi di campionamento in ingresso ed in uscita sia riconducibile ad un rapporto tra interi del tipo

$$\frac{T'}{T} = \frac{M}{L} \qquad \text{(D.2)}$$

Inoltre, data la linearità del filtraggio, è possibile considerare separatamente i due casi di interpolazione ($M = 1, L > 1$) e di decimazione ($M > 1, L = 1$). Il caso più generale { $M > 1, L > 1$ } è poi risolvibile come cascata di una interpolazione seguita da una decimazione.

Nel caso della decimazione, il ricampionamento si ottiene semplicemente eliminando M-1 campioni del segnale originario ogni M. Al fine di evitare aliasing, il segnale dovrebbe essere preventivamente limitato in banda tramite un filtro passa basso ideale con funzione di trasferimento

$$H(\omega) = \begin{cases} 1 & |\omega| \leq \frac{\pi}{M} \\ 0 & |\omega| > \frac{\pi}{M} \end{cases} \tag{D.3}$$

Indicando con h(n) la corrispondente risposta impulsiva, l'uscita si ottiene come

$$\begin{cases} w(n) = \sum_{k=-\infty}^{\infty} h(k)\ x(n-k) \\ y(m) = w(Mm) = \sum_{n=-\infty}^{\infty} h(n)\ x(mM-n) \end{cases} \tag{D.4}$$

Con una trasformazione di variabili, l'equazione precedente può essere riscritta come

$$y(m) = \sum_{n=-\infty}^{\infty} h(mM-n)\ x(n) \tag{D.5}$$

dalla quale si nota come i campioni derivanti dalla decimazione si ottengano traslando progressivamente la risposta impulsiva del filtro passa basso con un passo pari ad M campioni. Dal punto di vista della rappresentazione in frequenza, l'operazione di decimazione riespande lo spettro del segnale, limitato dal filtraggio nell'intervallo [$-\pi/M$, π/M], ad occupare tutto l'intervallo [$-\pi$, π].

Per eseguire, invece, un'interpolazione, è necessario introdurre L-1 campioni ogni due campioni adiacenti del segnale originale. Un modo per ottenere questo risultato è quello di considerare una versione sovracampionata

di un fattore L del segnale PAM corrispondente alla x(n), ottenuto introducendo dei campioni nulli nel segnale

$$w(m) = \begin{cases} x\left(\frac{M}{L}\right) & m = 0, \pm L, \pm 2L, \ldots \\ 0 & \text{altrimenti} \end{cases} \tag{D.6}$$

Dal punto di vista della rappresentazione in frequenza, ciò non altera lo spettro del segnale, coincidente con quello del corrispondente segnale PAM, il quale presenta repliche a partire dalla pulsazione $\omega = \pi/L$. Per passare allo spettro della x(t) sovracampionata con un fattore L è necessario eliminare tali repliche tramite un filtraggio passa basso, con frequenza di taglio corrispondente a $\omega = \pi/L$. Nel tempo, tale filtraggio si traduce in un'interpolazione che genera i campioni mancanti nella x(n). Per mantenere nel segnale sovracampionato l'ampiezza del segnale originario, il filtro di interpolazione deve avere un guadagno non unitario, ma pari ad L. Dato che, nel filtraggio, la risposta impulsiva h(n) si trova ad essere sistematicamente combinata con campioni nulli della w(n), la sommatoria di convoluzione può essere semplificata come

$$y(m) = \sum_{k=-\infty}^{\infty} w(k)\ h(m-k) = \sum_{n=-\infty}^{\infty} x(k)\ h(m-kL) \tag{D.7}$$

Analizzando quali campioni della h(n) vengono interessati dalla sommatoria, si nota che vengono utilizzati periodicamente L suoi sottoinsiemi g_m, tra loro scalati di un campione (fig. D.1). Tale sfasamento corrisponde ad un offset temporale che che è una frazione del periodo T pari a

$$\delta_m = \left[\frac{m}{L}\right]_{\text{mod } L} \tag{D.8}$$

È quindi possibile riscrivere l'equazione che fornisce la y(m) come l'uscita di un filtro tempo variante con risposta impulsiva g_m, tale che

$$\begin{cases} g_m = h[(n + \delta_{m)}\ T] \\ y(m) = \sum_{n=-\infty}^{\infty} g_m(n)\ x\left(\frac{m}{L} - n\right) \end{cases} \tag{D.9}$$

dove come indice della x deve considerarsi, ovviamente, un intero. Analogamente, è possibile considerare i campioni della y(m) come prodotti da un banco di L filtri in parallelo, ciascuno con una differente risposta impulsiva $g_m(n)$. L'uscita dell'interpolatore è poi ottenuto selezionando periodicamente l'uscita di uno degli L filtri che operano tra loro sfasati di un colpo di clock.

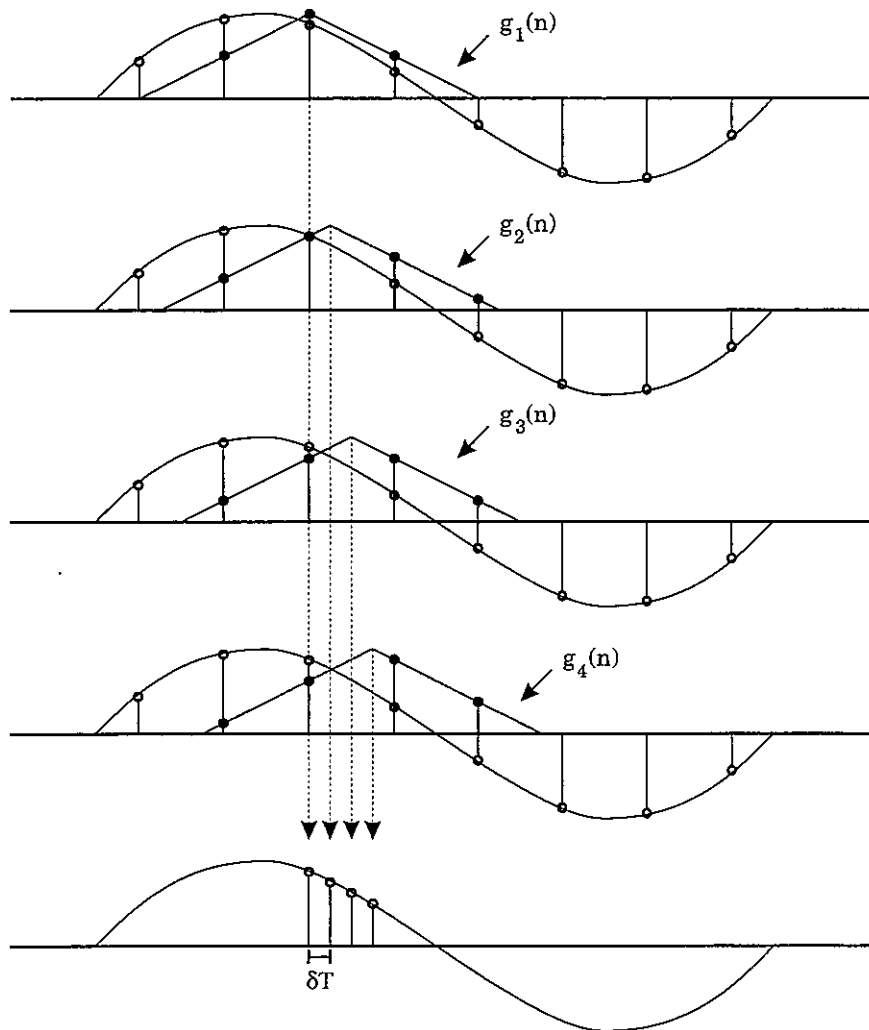



Fig. D.1 -Insieme di campioni della risposta impulsiva coinvolti in un'operazione di interpolazione.

Passando al problema dell'implementazione, è necessario considerare che i filtri utilizzati nella conversione di frequenza di campionamento sono tipicamente FIR a lunghezza finita N (fig. D.2), per cui l'uscita è ottenuta come

$$y(n) = \sum_{k=0}^{N-1} h(k)\ x(n-k) \tag{D.10}$$

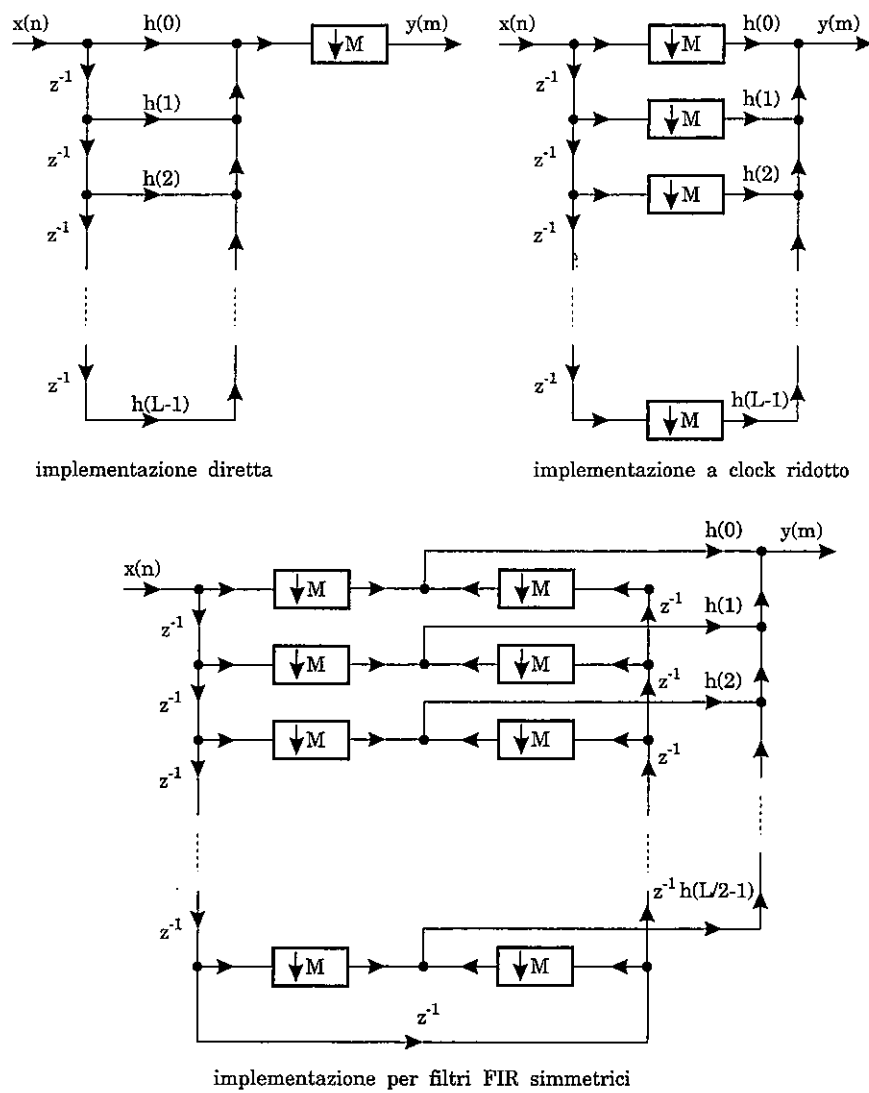



Fig. D.2 - Strutture di filtri FIR per decimazione.

Considerando per il momento il problema della decimazione, un inconveniente che si incontra implementando direttamente tale filtro è che esso lavora alla frequenza del segnale sovracampionato, mentre sarebbe auspicabile che lavorasse alla frequenza del segnale decimato. Per ottenere ciò è possibile anticipare l'operazione di decimazione prima del prodotto con i coefficienti del filtro come

$$y(m) = \sum_{k=0}^{N-1} h(k)\ x(M\,m - k) \tag{D.11}$$

ottenendo un filtro con clock pari a quello dell'uscita. Imponendo poi che i filtri siano a fase lineare e quindi con risposta impulsiva risulta simmetrica, la lunghezza del filtro può essere dimezzata in quanto

$$h(k) = h(N-1-k) \rightarrow y(n) = \sum_{k=0}^{N/2-1} h(k)\ \{\ x(n-k) + x[n-(N-1-k)]\ \} \tag{D.12}$$

Strutture analoghe a quelle presentate nel caso della decimazione possono essere realizzate nel caso di interpolazione. In ogni caso, volendo avere filtri sufficientemente selettivi, le loro dimensioni risultano essere notevoli. L'efficienza computazionale può essere incrementata riducendo la lunghezza del filtro tramite filtri polifase (fig. D.3). È stato mostrato come i campioni prodotti da un interpolatore possano essere pensati come l'uscita di uno di L filtri in parallelo, ciascuno con risposta impulsiva $g_m(n)$. Dato che la risposta impulsiva di ciascun ramo del filtro coincide con una versione decimata della h(n), che è diversa da zero solo negli istanti $L \times T$ ed è nulla altrimenti, è possibile eliminare i contributi di tali campioni nulli considerando un insieme di L filtri di dimensioni N/L tali che

$$p_k(n) = h(k + nL)\ ; \quad \begin{cases} k = 0, 1, ..., L-1 \\ n = 0, 1, ..., \frac{N}{L} - 1 \end{cases} \tag{D.13}$$

Ovviamente è opportuno che la lunghezza N del filtro prototipo sia multipla di L. Con una struttura polifase, per ogni campione dell'ingresso vengono prodotti L campioni d'uscita, uno per ogni ramo del filtro. L-1 di tali

campioni costituiscono l'interpolazione tra due campioni adiacenti dell'ingresso, necessari per il richiesto incremento della frequenza di campionamento.

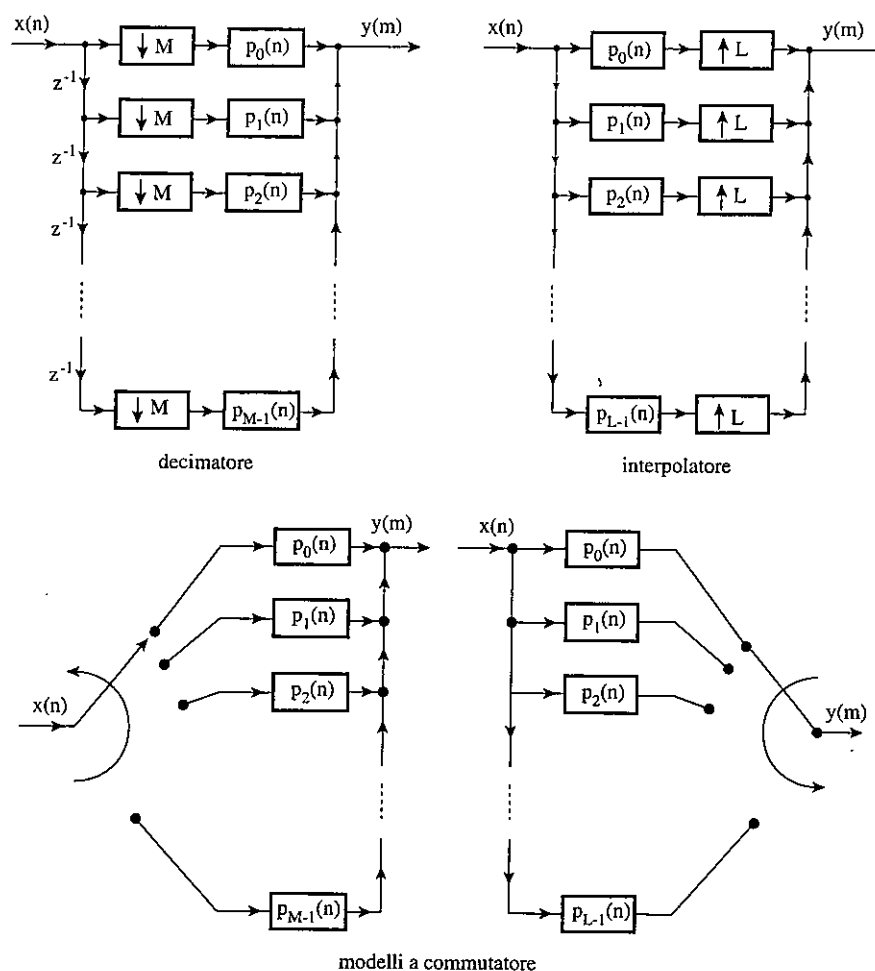



Fig. D3 - Strutture di filtri FIR polifase.

Per quanto riguarda la funzione di trasferimento di ciascuno degli L rami del filtro polifase, si ricorda che essi sono stati ottenuti come decimazione di un fattore L di un filtro passa basso ideale con frequenza di taglio corrispondente a $\omega = \pi/L$. Dato che la decimazione riespande lo spettro di un fattore L,

ciascun ramo del filtro polifase risulterà con funzione di trasferimento piatta nell'intervallo $[-\pi, \pi]$. L'unica differenza tra i differenti rami è quindi data dalla fase, da cui il nome del filtro.

Tipicamente la struttura di un filtro polifase non può essere semplificata ulteriormente sfruttando eventuali simmetrie. Infatti, pur utilizzando una h(n) simmetrica, le L risposte impulsive dei rami del filtro polifase non lo sono. D'altra parte, l'indipendenza tra i vari rami del filtro, ne rende possibile un'implementazione hardware parallela. Per quanto riguarda la decimazione, la relazione

$$y(m) = \sum_{k=-\infty}^{\infty} h(k)\ x(m\,M - k) \tag{D.14}$$

può essere riscritta come

$$y(m) = \sum_{k=0}^{M-1} \sum_{r=-\infty}^{\infty} h(r\,M + k)\ x[(m - r)\,M - k] \tag{D.15}$$

Definendo i coefficienti del filtro polifase di decimazione come

$$p_k^d(n) = h(nM + k)\ ;\quad k = 0, 1, ..., M-1 \tag{D.16}$$

si ottiene

$$y(m) = \sum_{k=0}^{M-1} \sum_{r=-\infty}^{\infty} p_k^d(r)\ x_k(m - r) \tag{D.17}$$

dove $x_k(n) = x(n\,M - k)$ è l'ingresso del ramo k-esimo del filtro. Di tali filtri è possibile dare un'interpretazione tramite un commutatore che preleva campioni nel caso della interpolazione e distribuisce campioni nel caso della decimazione ruotando in senso antiorario. Ciò vuol dire, ad esempio, che nella decimazione il primo campione ingresso viene assegnato al ramo di ordine maggiore e così via. È possibile realizzare strutture con senso di rotazione del commutatore orario definendo i filtri polifase come

$$\begin{cases} p_k^i(n) = h(nL - k)\ ;\quad k = 0, 1, ..., L-1 \\ p_k^d(n) = h(nM - k)\ ;\quad k = 0, 1, ..., M-1 \end{cases} \tag{D.18}$$

Un'applicazione importante di multirate DSP si ha nell'elaborazione di segnali passa banda, cioè di segnali con una banda limitata ω_Δ, ma centrati in $\omega_0 \neq 0$ (fig. D.4). Tale esigenza si incontra, ad esempio, in un banco di filtri: il fine è quello di ricavare il corrispondente segnale in banda base X(m), cioè il segnale che, a parità di banda, risulta centrato nell'origine. Ciò è concettualmente ottenibile innanzitutto traslando il segnale passa banda nell'origine tramite una modulazione e quindi decimandolo dopo il necessario filtraggio antialiasing tramite un passa basso con frequenza di taglio corrispondente a $\omega_\Delta/2$. Indicando con h(n) la risposta impulsiva di tale filtro, si ottiene

$$X(m) = \sum_{n=-\infty}^{\infty} h(mM - n) \left[x(n)\ e^{-jn\omega_0} \right] \tag{D.19}$$

dove M è il fattore di decimazione che si ricava dal rapporto tra la frequenza centrale del segnale e la sua banda. Per ricostruire il segnale originale dalla sua versione decimata, invece, è innanzitutto necessario interpolarlo per poi modularlo, riportandolo su ω_0. Indicandolo con f(n) la risposta impulsiva del filtro di interpolazione, si ottiene

$$x(n) = e^{jn\omega_0} \left[\sum_{m=-\infty}^{\infty} X_k(m)\ f(n - mM) \right] \tag{D.20}$$

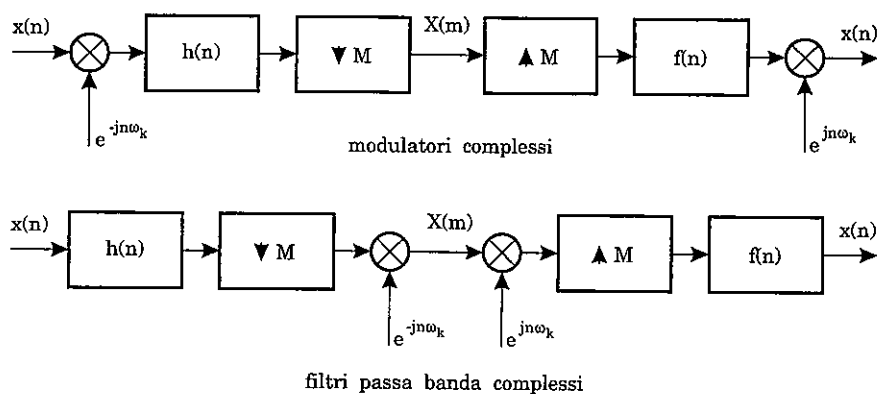



Fig. D.4 - Differenti modulazioni per segnali passabanda.

Dato che il segnale d'uscita risulta essere complessa, a causa del prodotto con gli esponenziali, tale tecnica è detta di modulazione complessa (fig. D.5). La struttura del sistema di decimazione/interpolazione può essere anche differente da quella presentata, portando le modulazioni come ultima operazione nella decimazione e come prima della interpolazione. Con tale soluzione, però, è necessario l'uso di filtri passa banda centráti sul segnale stesso, al posto dei passa basso del caso precedente.

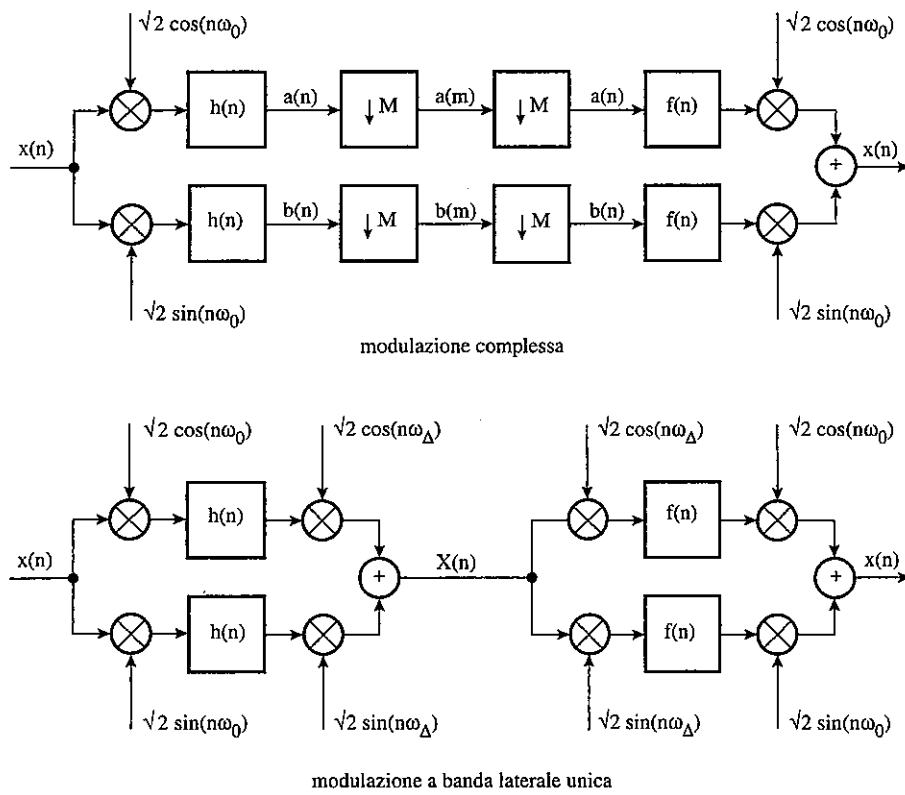



Fig. D.5 - Modulazione complessa ed a banda laterale unica

Analiticamente l'espressione del decimatore può essere ottenuta dalla precedente moltiplicandola per $e^{jmM\omega_0} \cdot e^{-jmM\omega_0}$ e riarrangiando i termini come

$$X(m) = e^{-jmM\omega_0} \sum_{n=-\infty}^{\infty} \left[h(mM-n)\; e^{j(mM-n)\omega_0} \right] x(n) \qquad \text{(D.21)}$$

dove il termine tra parentesi quadre corrisponde alla risposta impulsiva del filtro passa banda di ampiezza ω_Δ centrato su ω_0. Analogamente, per l'interpolazione

$$x(n) = \sum_{m=-\infty}^{\infty} \left[X(m)\, e^{jmM\omega_0} \right] \left[f(n - mM)\, e^{j(n-mM)\omega_0} \right] \tag{D.22}$$

Le relazioni presentate forniscono un mappaggio del segnale passa banda in un segnale complesso in banda base nell'intervallo $[-\omega_\Delta/2, \omega_\Delta/2]$ e viceversa. Per applicazioni di codifica del segnale, invece, è più interessante considerare un mappaggio nell'intervallo $[0, \omega_\Delta]$, in grado di fornire un segnale reale. Per tale mappaggio, noto come modulazione SSB per l'analogia con le analoghe tecniche trasmissive, al segnale ottenuto dalla modulazione complessa viene applicata una traslazione in frequenza di metà larghezza di banda, tramite il prodotto con $e^{jn\frac{\omega_\Delta}{2}}$.

D.2 BANCHI DI FILTRI POLIFASE

La funzione di un banco di filtri può essere sia quella di separare un segnale nelle componenti che fanno capo a differenti bande spettrali (filtro di analisi), sia quella di ricombinare i contributi di bande differenti in un unico segnale (filtro di sintesi) (fig. D.6). Per tali operazioni è necessario considerare i differenti segnali passa banda relativi a ciascuna sottobanda, applicando ad essi i risultati ottenuti in Appendice D.1.

Per quanto ci riguarda, il caso di maggiore interesse è quello di un banco di filtri in grado di generare K bande uniformemente spaziate. In tal caso, la sottobanda relativa alle frequenze inferiori risulta modulata SSB. Tale disposizione delle sottobande è normalmente indicata come even-stacking, a differenza del caso in cui, pur avendo bande uniformi, la prima sottobanda è centrata nell'origine, nel qual caso si parla di struttura odd-stacking del banco di filtri.

Considerando modulazioni complesse, il segnale relativo alla banda k-esima $X_k(m)$ è quindi ottenibile come

$$\begin{cases} X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)\, x(n)\, e^{-jn\omega_k} \\ \omega_k = \dfrac{2k\pi}{K} \end{cases} \tag{D.23}$$

mentre per il filtro di sintesi, indicando con $\hat{X}_k$ il risultato dell'elaborazione eseguita sul segnale relativo alla banda k-esima, si ricava

$$\hat{x}(n) = \frac{1}{K}\sum_{k=0}^{K-1} \hat{x}_k(n) = \frac{1}{K}\sum_{k=0}^{K-1} e^{jn\omega_k} \left[\sum_{m=-\infty}^{\infty} \hat{X}_k(m)\, f(n-nM) \right] \qquad (D.24)$$

dove h(m) è la risposta impulsiva del filtro passa basso di analisi e f(m) quella utilizzata nella sintesi. Riscrivendo il filtro di sintesi come

$$\hat{x}(n) = \sum_{m=-\infty}^{\infty} f(n-nM) \left[\frac{1}{K}\sum_{k=0}^{K-1} e^{jn\omega_k}\, \hat{X}_k(m) \right] \qquad (D.25)$$

si vede come le uscite dei banchi di analisi e sintesi siano ottenibili tramite trasformata discreta di Fourier, da cui il nome di DFT filter banks utilizzati nel caso di banchi con bande uniformemente spaziate. Il massimo valore ammissibile per il fattore di decimazione M coincide con il numero di sottobande K, nel qual caso si parla si bande critically-sampled.

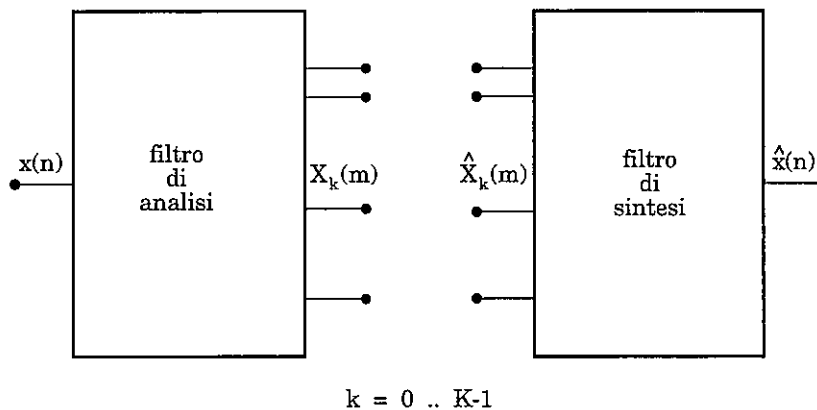



Fig. D.6 - Struttura del sistema di analisi e sintesi.

Utilizzando filtri passa banda complessi la struttura cambia, con le modulazioni che diventano l'ultima operazione nell'analisi e la prima della sintesi ed ottenendo

$$X_k(m) = e^{-jmM\omega_k} \sum_{n=-\infty}^{\infty} \left[h(mM-n)\ e^{j(mM-n)\omega_k} \right] x(n)$$

$$x(n) = \frac{1}{K} \sum_{k=0}^{K-1} \left\{ \sum_{m=-\infty}^{\infty} \left[X_k(m)\ e^{jmM\omega_k} \right] \left[f(n-mM)\ e^{j(n-mM)\omega_k} \right] \right\} \qquad (D.26)$$

Nel caso di K bande uniformi, quindi, le funzioni di trasferimento di ciascun filtro passa banda sono derivate per traslazione di un unica funzione prototipo passa basso $H(\omega)$, in modo tale che per la banda k-esima si abbia

$$\begin{cases} H_k(\omega) = H(\omega - \omega_k) \\ h_k(n) = h(n) \cdot e^{jn\omega_k} \\ \omega_k = \dfrac{2k\pi}{K} \end{cases} \qquad (D.27)$$

Un miglioramento dal punto di vista dell'efficienza computazionale si ha utilizzando filtri polifase (fig. D7). In tal caso, la risposta impulsiva del ramo di ordine ρ per la banda k-esima ed il suo ingresso sono pari a

$$\begin{cases} p_{\rho,k}(m) = h_k(mM-\rho) \\ x_\rho(m) = x(mM-\rho) \end{cases} \quad ; \quad \rho = 1, \ldots, K-1 \qquad (D.28)$$

Secondo tale impostazione sembrerebbe che i filtri polifase da utilizzare sarebbero in numero pari al prodotto del numero di sottobande per il fattore di decimazione. Fortunatamente, però, sfruttando la relazione

$$h_k(n) = h(n) \cdot e^{jn\frac{2k\pi}{K}} = h(n) \cdot e^{jn\frac{2k\pi}{M}} \qquad (D.29)$$

si ricava

$$p_{\rho,k}(m) = h(mM-\rho) \cdot e^{j(mM-\rho)\frac{2k\pi}{M}} = h(mM-\rho) \cdot e^{-j\frac{2k\rho\pi}{M}} \qquad (D.30)$$

Se si definisce il ρ-esimo ramo dell'equivalente filtro polifase passa basso come

$$p_\rho(m) = h(mM - \rho) \tag{D.31}$$

allora è possibile riscrivere il filtro polifase della banda k-esima come

$$p_{\rho,k}(m) = p_\rho(m) \cdot e^{-j\frac{2k\rho\pi}{M}} \tag{D.32}$$

Sostituendo nelle equazioni precedenti ed indicando con $q_\rho(m)$ il ρ-esimo ramo del polifase passa basso equivalente del filtro di sintesi, si ottiene

$$\begin{cases} X_k(m) = \sum_{\rho=0}^{M-1} \left[\sum_{r=-\infty}^{\infty} p_\rho(r)\, x_\rho(m-r) \right] e^{-j\frac{2k\rho\pi}{M}} \\ \hat{x}(n) = \sum_{m=-\infty}^{\infty} q_\rho(n-m) \left[\frac{1}{M} \sum_{k=0}^{M-1} \hat{X}_k(m)\, e^{j\frac{2k\rho\pi}{M}} \right] \end{cases} \tag{D.33}$$

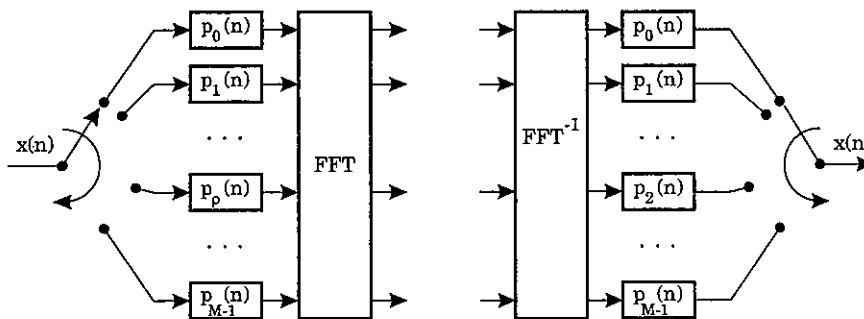



Fig. D.7 - Banco di filtri polifase.

Dall'analisi di tali relazioni, si nota come l'efficienza computazionale nasca, innanzitutto, dal poter condividere il filtro $p_\rho(m)$ tra le differenti bande, Inoltre, si nota come l'uscita del filtro di analisi sia ottenibile come DFT delle uscite dei singoli filtri polifase e che il filtro di sintesi si ottiene applicando i filtri polifase alla DFT inversa degli X_k. Un ulteriore incremento di efficienza computazionale nasce, quindi, dall'uso di FFT.

Un problema che si incontra nel filtraggio passa banda è la presenza di aliasing, introdotto dalla decimazione, a seguito dell'uso di filtri non ideali. In tal modo l'energia in ciascuna sottobanda è influenzata dal contributo delle bande adiacenti. Questo fenomeno è tanto più sensibile per quelle bande a bassa energia, per le quali il contributo dovuto alla non idealità dei filtri può essere predominante. La realizzazione ottimale del banco dei filtri è, quindi, un punto chiave e la scelta dei filtri h(n) ed f(n) deve essere tale da minimizzare l'aliasing. È possibile analizzare tale fenomeno sia nel dominio del tempo o in frequenza.

Analizzando l'aliasing nel dominio del tempo, si vuole vedere sotto quali condizioni ne è possibile l'eliminazione. Imporre che $\hat{x}(n) = x(n)$ è possibile sostituendo l'espressione del filtro di analisi in quello di sintesi, ottenendo

$$\begin{aligned}\hat{x}(n) &= \sum_{m=-\infty}^{\infty} f(n-nM)\left[\frac{1}{K}\sum_{k=0}^{K-1} e^{jn\omega_k} \sum_{r=-\infty}^{\infty} h(mM-r)\; x(r)\; e^{-jr\omega_k}\right] \\ &= \sum_{m=-\infty}^{\infty}\sum_{r=-\infty}^{\infty} x(r)\; f(n-nM)\; h(mM-r)\left[\frac{1}{K}\sum_{k=0}^{K-1} e^{j(n-r)\omega_k}\right]\end{aligned} \qquad (D.34)$$

Dato che l'esponenziale è pari ad 1 per $r = n - sK$, con s intero, l'espressione precedente può essere riscritta come

$$\hat{x}(n) = \sum_{s=-\infty}^{\infty} x(n-sK) \sum_{m=-\infty}^{\infty} f(n-nM)\; h(mM-n+sK) \qquad (D.35)$$

affinché sia valida l'uguaglianza $\hat{x}(n) = x(n)$, la seconda sommatoria deve annullarsi per $s \neq 0$ e deve valere uno per $s = 0$. Indicando le versioni traslate delle risposte impulsive come

$$\begin{cases} h_m(n) = h(n-mM) \\ f_m(n) = f(n-mM) \end{cases} \qquad (D.36)$$

la relazione precedente si riscrive come

$$\hat{x}(n) = x(n)\sum_{m=-\infty}^{\infty} f_m(n)\; h_m(-n) + \sum_{s=-\infty, s\neq 0}^{\infty} x(n-sK) \sum_{m=-\infty}^{\infty} f_m(n)\; h_m(sK-n) \qquad (D.37)$$

Questa equazione ci permette di definire quali siano le condizioni di perfetta ricostruzione del segnale. Affinché sia valida l'uguaglianza $\hat{x}(n) = x(n)$,dalla prima sommatoria si ricava che

$$\sum_{n=-\infty}^{\infty} f_m(n)\, h_m(-n) = 1 \tag{D.38}$$

il che indica che la somma del prodotto delle risposte impulsive traslate deve essere unitaria per qualsiasi valore di n. Dalla seconda sommatoria, che è necessario annullare, si comprende come sia necessario annullare le infinite repliche poste ad intervalli di K campioni, introdotte dalla decimazione. Ciò è ottenibile imponendo

$$\sum_{m=-\infty}^{\infty} f_m(n)\; h_m(sK - n) = 0\,, \quad s \neq 0 \tag{D.39}$$

il che si raggiunge solo con filtri di lunghezza inferiore a K. Dato che la lunghezza dei filtri risultante da tale vincolo sarebbe estremamente modesta, essi non risulterebbero soddisfacenti dal punto di vista dell'analisi in frequenza in quanto poco selettivi. Passando, infatti, all'analisi in frequenza, è conveniente considerare un modello a filtri passa banda complessi, semplificato eliminando le modulazioni terminali (che si elidono) e sostituendo la decimazione/interpolazione con un modulatore che moltiplica il segnale per la sequenza

$$s(n) = \begin{cases} 1 & n = \pm\, mM\,,\ \text{m intero} \\ 0 & \text{altrimenti} \end{cases} \tag{D.40}$$

La rappresentazione in frequenza del legame ingresso/uscita che si ottiene è pari a

$$\hat{X}(\omega) = X(\omega)\frac{1}{K}\sum_{k=0}^{K-1} F_k(\omega)\, H_k(\omega) + \sum_{l=0}^{M-1} X\!\left(\omega - \frac{2l\pi}{M}\right)\frac{1}{K}\sum_{k=0}^{K-1} F_k(\omega)\, H_k\!\left(\omega - \frac{2l\pi}{M}\right) \tag{D.41}$$

dove

$$\begin{cases} H_k(\omega) = H(\omega - \omega_k) \\ F_k(\omega) = F(\omega - \omega_k) \end{cases} \tag{D.42}$$

sono i filtri di analisi e sintesi della banda k-esima. Mentre il primo termine di tale equazione implica che la funzione di trasferimento complessiva sia unitaria, il secondo impone funzioni passa basso ideali, approssimabili solo con filtri di dimensione elevata. In conclusione, non è possibile raggiungere completamente la cancellazione dell'aliasing né nel dominio del tempo né in frequenza.

Tutte le relazioni precedenti sono state ricavate per una struttura even-stacking del banco di filtri. Per un'organizzazione odd-stacking deve risultare

$$\omega_k = \frac{2\pi}{K}\left(k + \frac{1}{2}\right) \tag{D.43}$$

Si è dimostrato come sia il filtro di analisi che quello di sintesi sono legati ad un mappaggio tramite DFT dal dominio del tempo alla frequenza e viceversa. Il mappaggio tramite DFT, però, è sempre tale da legate l'origine temporale $n = 0$ a quella in frequenza $\omega = 0$. Volendo permettere il mappaggio tra due qualsiasi origini nei due domini, è possibile definire una trasformata discreta di Fourier generalizzata come

$$\begin{cases} X_k^{GDFT} = \sum_{n=0}^{K-1} x(n)\, e^{-j(k+k_0)(n+n_0)\frac{2\pi}{K}} \\ x(n) = \frac{1}{K} \sum_{n=0}^{K-1} X^{GFDT}\, e^{j(k+k_0)(n+n_0)\frac{2\pi}{K}} \end{cases} \tag{D.44}$$

dove n_0 e k_0 sono le nuove origini nel tempo ed in frequenza. Applicando tale generalizzazione a banchi di filtri di analisi e sintési che, nel caso di struttura odd-stacking, sono caratterizzati $k_0 = 1/2$, si ottiene

$$\begin{cases} X_k^{GDFT}(m) = \sum_{n=-\infty}^{\infty} h(mM - n)\, x(n)\, e^{-j(n+n_0)\left(k+\frac{1}{2}\right)\frac{2\pi}{K}} \\ x(n) = \sum_{m=-\infty}^{\infty} f(n - mM)\, \frac{1}{K} \sum_{k=0}^{K-1} X_k^{GDFT}(m)\, e^{j(n+n_0)\left(k+\frac{1}{2}\right)\frac{2\pi}{K}} \end{cases} \tag{D.45}$$

Anche per tale banco, la struttura può evolvere verso strutture con filtri passa banda complessi o polifase. Inoltre, al fine di ottenere segnali reali, è possibile trasformarlo applicando una modulazione SSB

$$X_k^{SBB}(m) = \mathrm{Re}\left[X_k^{GDFT}(m)\, e^{j\frac{mM}{2}\omega_\Delta}\right] \tag{D.46}$$

da cui, posto M = K/2, deriva

$$\begin{aligned} X_k(m) &= \cos\left(\frac{m\pi}{2}\right)\sum_{n=-\infty}^{\infty} x(n)\cdot h\left(P-1+\frac{mK}{2}-n\right)\cdot\cos\left[\frac{2\pi}{K}\cdot\left(k+\frac{1}{2}\right)\cdot(n+n_0)\right] \\ &+ \sin\left(\frac{m\pi}{2}\right)\sum_{n=-\infty}^{\infty} x(n)\cdot h\left(P-1+\frac{mK}{2}-n\right)\cdot\sin\left[\frac{2\pi}{K}\cdot\left(k+\frac{1}{2}\right)\cdot(n+n_0)\right] \end{aligned} \tag{D.47}$$

Tale relazione può essere formalmente semplificata in

$$\begin{cases} X_k(m) = \cos(\pi mk)\displaystyle\sum_{n=-\infty}^{\infty} x(n)\cdot h(mM-n)\cdot\cos\left[\frac{2\pi}{K}\left(k+\frac{1}{2}\right)(mM-n-n_0)\right] \\ \hat{x}(n) = \displaystyle\sum_{m=-\infty}^{\infty}\sum_{k=0}^{\frac{K}{2}-1} \cos(\pi mk)\,\hat{X}_k(m)\cdot f(n-mM)\cdot\cos\left[\frac{2\pi}{K}\left(k+\frac{1}{2}\right)(n-mM+n_0)\right] \end{cases} \tag{D.48}$$

D.3 BANCHI DI FILTRI QMF

Una tecnica differente per la realizzazione del banco è quella dei Quadrature-Mirror Filters (QMF) (fig. D.8). Un banco QMF permette la divisione dello spettro del segnale in due sotto-bande. Ciò è ottenuto tramite un filtraggio passa-alto abbinato ad un filtraggio complementare passa-basso. Banchi con un numero maggiore di due sottobande si ottengono ripetendo il filtraggio su entrambi (per sottobande uniformi) o uno solo dei segnali d'uscita (fig. D.9). La struttura risultante è ad albero, bilanciato o meno. Al fine di minimizzare l'effetto dell'aliasing, la somma delle due funzioni di trasferimento deve essere piatta. Se si indica con $H_1(e^{j\omega T})$ e $H_2(e^{j\omega T})$ la

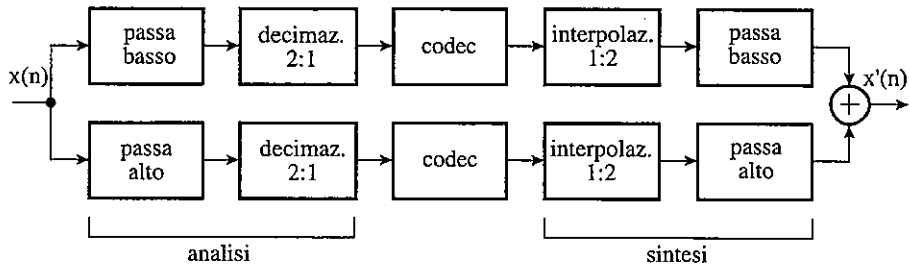



Fig. D.8 - Diagramma di un banco QMF.

funzione di trasferimento del filtro passa alto e passa basso di analisi, i segnali da loro prodotti sono pari a

$$X_0(\omega) = \frac{1}{2}\left[X\left(\frac{\omega}{2}\right)H_0\left(\frac{\omega}{2}\right) + X\left(\frac{\omega+2\pi}{2}\right)H_0\left(\frac{\omega+2\pi}{2}\right)\right]$$

$$X_1(\omega) = \frac{1}{2}\left[X\left(\frac{\omega}{2}\right)H_1\left(\frac{\omega}{2}\right) + X\left(\frac{\omega+2\pi}{2}\right)H_1\left(\frac{\omega+2\pi}{2}\right)\right] \quad \text{(D.49)}$$

Indicando con $F_0(\omega)$ e $F_1(\omega)$ la funzione di trasferimento del filtro passa basso e passa alto di sintesi, in fase di ricostruzione si ottiene

$$\hat{X}(\omega) = X_0(2\omega)\,F_0(\omega) + X_1(2\omega)\,F_1(\omega) \quad \text{(D.50)}$$

Sostituendo in tale espressione quella dei segnali prodotti da ciascuna sottobanda si ottiene

$$\hat{X}(\omega) = \frac{1}{2}\left[H_0(\omega)\,F_0(\omega) + H_1(\omega)\,F_1(\omega) \right] X(\omega) \; +$$

$$+ \; \frac{1}{2}\left[H_0(\omega+\pi)\,F_0(\omega) + H_1(\omega+\pi)\,F_1(\omega) \right] X(\omega+\pi) \quad \text{(D.51)}$$

Imponendo che si annulli l'effetto dell'aliasing è necessario annullare il secondo termine della precedente equazione, ottenendo

$$H_0(\omega+\pi)\,F_0(\omega) + H_1(\omega+\pi)\,F_1(\omega) = 0 \quad \text{(D.52)}$$

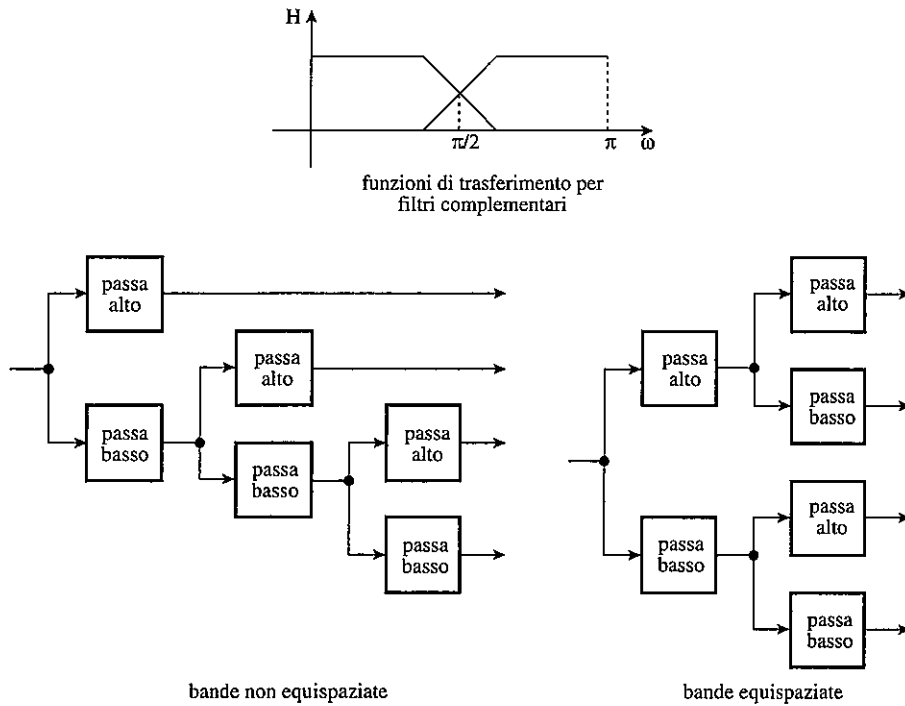



Fig. D.9 - Filtraggio passa banda tramite cascata di filtri complementari.

Per trovare le condizioni tali da soddisfare tale vincolo è necessario ridurre le variabili del problema. A tal fine si imporre innanzitutto che il filtro passa altro sia una replica traslata del passa basso di analisi

$$\begin{cases} H_0(\omega) = H(\omega) \\ H_1(\omega) = H(\omega + \pi) \end{cases} \qquad (D.53)$$

da cui deriva

$$H(\omega + \pi)\, F_0(\omega) + H(\omega)\, F_1(\omega) = 0 \;\rightarrow\; \begin{cases} F_0(\omega) = 2\, H(\omega) \\ F_1(\omega) = 2\, H(\omega + \pi) \end{cases} \qquad (D.54)$$

dove il fattore di scala 2 serve per rendere unitaria la funzione di trasferimento complessiva. Infatti

$$\hat{X}(\omega) = [\, H^2(\omega) - H^2(\omega + \pi)\,]\, X(\omega) \qquad (D.55)$$

L'esser riusciti ad eliminare il secondo termine dell'equazione precedente vuol dire che l'aliasing introdotto nel filtro di sintesi da parte della decimazione è cancellato dalle repliche introdotte dall'interpolazione nel filtro di sintesi. Nel dominio del tempo, le risposte impulsive dei filtri sono date da

$$\begin{cases} h_0(n) = h(n) \\ h_1(n) = (-1)^n\, h(n) \end{cases} ; \quad \begin{cases} g_0(n) = 2\, h(n) \\ g_1(n) = -2\,(-1)^n\, h(n) \end{cases} \qquad (D.56)$$

Rimane una sola incognita, data da $H(\omega)$, per la determinazione della quale si possono seguire differenti criteri [Cro83]. Una classe di filtri particolarmente diffusa è quella dei filtri FIR a fase lineare. Il vantaggio dell'assenza di distorsioni di fase risulta particolarmente utile nel caso di connessioni in cascata, eliminando la necessità di equalizzazioni ad ogni stadio. La risposta impulsiva di un filtro FIR a fase lineare risulta essere simmetrica, per cui

$$h(n) = h(N-1-n)\,; \quad n = 0, 1, ..., N-1 \qquad (D.57)$$

e la funzione di trasferimento è data da una funzione reale ed un termine di fase lineare

$$H(\omega) = H_r(\omega)\, e^{-j\omega \frac{N-1}{2}} \qquad (D.58)$$

La funzione di trasferimento complessiva nel caso di N pari e dispari risulta

$$\begin{cases} \hat{X}(\omega) = \left[\, |H(\omega)|^2 + |H(\omega+\pi)|^2 \right] e^{j\omega(N-1)}\, X(\omega)\,; & \text{N pari} \\ \hat{X}(\omega) = \left[\, |H(\omega)|^2 - |H(\omega+\pi)|^2 \right] e^{j\omega(N-1)}\, X(\omega)\,; & \text{N dispari} \end{cases}$$

$$A(\omega) = |H(\omega)|^2 + |H(\omega+\pi)|^2 = 1 \qquad (D.59)$$

Dato che con N dispari, la funzione di trasferimento si annulla per $\omega = \pi/2$ in quanto $H(\pi/2) = H(3\pi/2)$, si utilizzano solamente filtri con N pari. Fissato l'ordine del filtro, l'unico filtro che rende unitaria la funzione di trasferimento complessiva è il filtro

$$|H(\omega)|^2 = \cos^2\alpha\omega \qquad (D.60)$$

le cui caratteristiche selettive, però, sono povere. Ogni altro filtro introduce delle distorsioni di ampiezza. Per la definizione di differenti funzioni di trasferimento si ricorre a procedure iterative per la contemporanea minimizzazione della distorsione introdotta dal filtro e dell'energia in banda soppressa, cioè dell'errore quadratico

$$\varepsilon = w\int_{\omega_s}^{\pi} |H(\omega)|^2 \, d\omega + (1-w)\int_0^{\frac{\pi}{2}} [A(\omega)-1]^2 \, d\omega \, ; \quad 0 \leq w \leq 1 \tag{D.61}$$

il cui primo termine rappresenta l'energia in banda soppressa, il secondo la distorsione in banda e w un coefficiente di pesatura.

Appendice E

RICHIAMI SU TRASFORMATE NUMERICHE

E.1 TRASFORMATA DI FOURIER DISCRETA

Data una funzione x(t) continua in $-\infty < t < \infty$, la sua trasformata ed antitrasformata di Fourier è data da

$$X(\omega) = F[x(t)] \equiv \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, dt$$

$$x(t) = F_{-1}[X(\omega)] \equiv \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\infty} X(\omega)\, e^{j\omega t}\, dtdté \qquad (E.1)$$

La trasformata di Fourier può essere estesa dall'analisi dei segnali continui ai segnali a tempo-discreto (Discrete Time Fourier Transform: DTFT) tramite la sostituzione della variabile continua t con l'indice n, ottenendo

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)\, e^{-j\omega n} \qquad (E.2)$$

La trasformata di Fourier rappresenta la sequenza x(n) come sovrapposizione di esponenziali complessi $e^{j\omega n}$ di ampiezza infinitesima

$$\frac{X(e^{j\omega})\, d\omega}{2\pi} \qquad (E.3)$$

ed è quindi una funzione complessa continua. Inoltre, data la periodicità dell'esponenziale complesso, la DTFT risulta essere una funzione periodica in

frequenza. Ciò si ripercuote nell'antitrasformata, nella quale gli estremi di integrazione dell'integrale di definizione sono limitati in un intervallo di ampiezza pari a 2p

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})\, e^{j\omega n}\, d\omega \tag{E.4}$$

Es.: si consideri la DTFT di una rect(n) discreta

$$x(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{altrove} \end{cases} \tag{E.5}$$

la sua DTFT è data da

$$X\left(e^{j\omega}\right) = \sum_{n=0}^{N-1} e^{-jn\omega} = \frac{1 - e^{-jN\omega}}{1 - e^{-j\omega}} = \frac{\sin(^{N\omega}/_2)}{\sin(^{\omega}/_2)}\, e^{-j(N-1)\frac{\omega}{2}} \tag{E.6}$$

Il modulo della DTFT è dato dal solo rapporto tra sinusoidi, che è una funzione periodica, il cui primo zero è in $\omega = \frac{2\pi}{N}$. L'esponenziale è relativo alla sola traslazione della rect(n), che non risulta centrata nell'origine. È interessante confrontare questa DTFT con la trasformata di Fourier di un'analoga rect(t) continua, data da

$$X\left(e^{j\omega}\right) = \int_{-\frac{T}{2}}^{\frac{T}{2}} A\, e^{-j\omega t}\, dt = A\,\frac{\sin(^{T\omega}/_2)}{^{\omega}/_2} \tag{E.7}$$

cioè una sinc(ω) il cui primo zero è in $\omega = \frac{2\pi}{T}$. Diagrammando le due funzioni in modo che gli zeri coincidano (che simula l'ottenimento della rect(n) dalla rect(t) campionata con N punti) si nota l'effetto dell'aliasing che fa discostare la DTFT dalla FT tanto più quanto ci si allontana dall'origine (fig. E.1).

Come la DTFT è l'analoga della trasformata, data una sequenza di N elementi periodica con periodo T, o una sequenza di durata finita periodicizzata, è possibile introdurre nel caso discreto l'equivalente della serie di Fourier (Discrete Fourier Series: DFS) ripetendo la sostituzione della variabile continua t con l'indice n. A causa della periodicità delle sequenze sinusoidali,

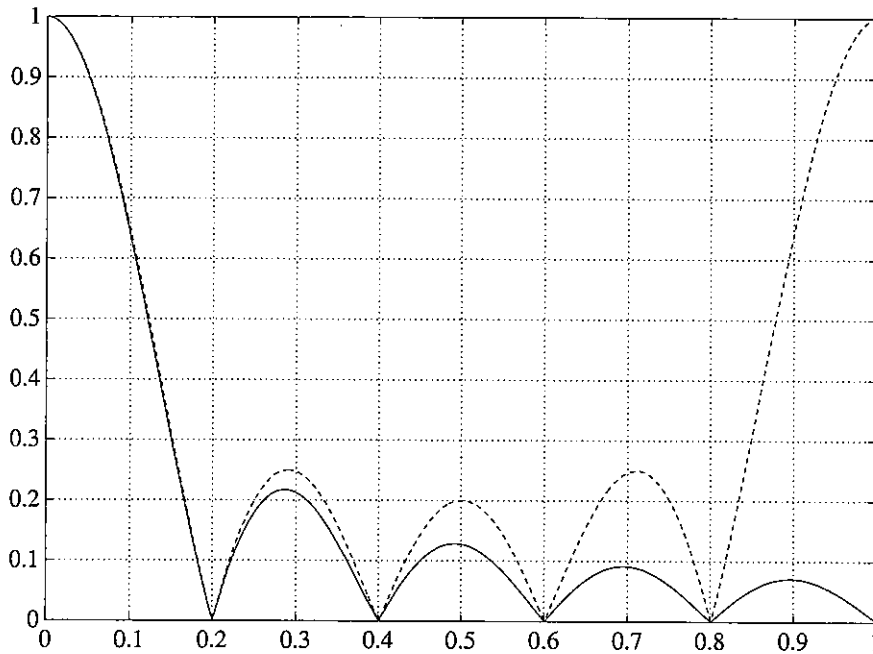



Fig. E.1 - FT e DTFT della rect.

l'analisi in frequenza ha senso solamente nell'intervallo limitato $[0 \leq \omega < 2\pi]$. Per sequenze con pulsazione ω_0 rappresentate da N punti, inoltre, si possono considerare solamente le prime N armoniche, per cui gli indici delle sommatorie nelle definizioni della serie vengono limitati ad N.

$$x(mT_0) = \sum_{k=0}^{N-1} X(k)\, e^{jkm\left(\frac{2\pi}{N}\right)} = \sum_{k=0}^{N-1} X(k)\, W_N^{-km}$$

$$X(k) = \frac{1}{N}\sum_{m=0}^{N-1} x(mT_0)\, e^{jkm\left(\frac{2\pi}{N}\right)} = \frac{1}{N}\sum_{m=0}^{N-1} x(mT_0)\, W_N^{km} \tag{E.8}$$

dove $T_0 = \frac{T}{N}$ e $W_N = e^{-j\left(\frac{2\pi}{N}\right)}$ è radice n-esima dell'unità interpretabile come campionamento di funzioni trigonometriche. La limitazione in frequenza comporta un effetto di aliasing che, come per la DTFT, fa si che i coefficienti della serie siano differenti dagli analoghi dei sistemi continui. Il calcolo numerico

della serie di Fourier è ricondotto alla soluzione del sistema di N equazioni espresse dalla definizione, una per ogni valore di m.

Esistendo una relazione diretta tra coefficienti della serie e valori dello spettro, le equazioni precedenti stabiliscono una corrispondenza tra N campioni di una sequenza x(n) e N campioni di una trasformata X(k), per cui la serie di Fourier per segnali tempo discreto è anche interpretabile come trasformata discreta di Fourier (DFT), definita come:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)\, W_N^{-kn}$$

$$X(k) = \sum_{k=0}^{N-1} x(n)\, W_N^{kn} \qquad \text{(E.9)}$$

Es.: data la serie x(n) = {0, 2, 1, 1}, la sua DFT è data da

$$W_4^{kn} = e^{-j\frac{2\pi}{4}kn} = \cos\left(\frac{2\pi}{4}kn\right) - j\sin\left(\frac{2\pi}{4}kn\right)$$

$$W_4^0 = \cos(0) - j\sin(0) = 1$$

$$W_4^1 = \cos\left(\frac{\pi}{2}\right) - j\sin\left(\frac{\pi}{2}\right) = -j$$

$$W_4^2 = \cos(\pi) - j\sin(\pi) = -1$$

$$W_4^3 = \cos\left(\frac{3}{2}\pi\right) - j\sin\left(\frac{3}{2}\pi\right) = +j$$

$$X(0) = x(0)\, W_4^0 + x(1)\, W_4^0 + x(2)\, W_4^0 + x(3)\, W_4^0 = 4$$

$$X(1) = x(0)\, W_4^0 + x(1)\, W_4^1 + x(2)\, W_4^2 + x(3)\, W_4^3 = 1 - j$$

$$X(2) = x(0)\, W_4^0 + x(1)\, W_4^2 + x(2)\, W_4^0 + x(3)\, W_4^2 = 2$$

$$X(3) = x(0)\, W_4^0 + x(1)\, W_4^3 + x(2)\, W_4^2 + x(3)\, W_4^1 = 1 + j \qquad \text{(E.10)}$$

È possibile ora analizzare il legame esistente tra lo spettro di una funzione continua e la DFT di una sequenza finita, ottenuta valutando la

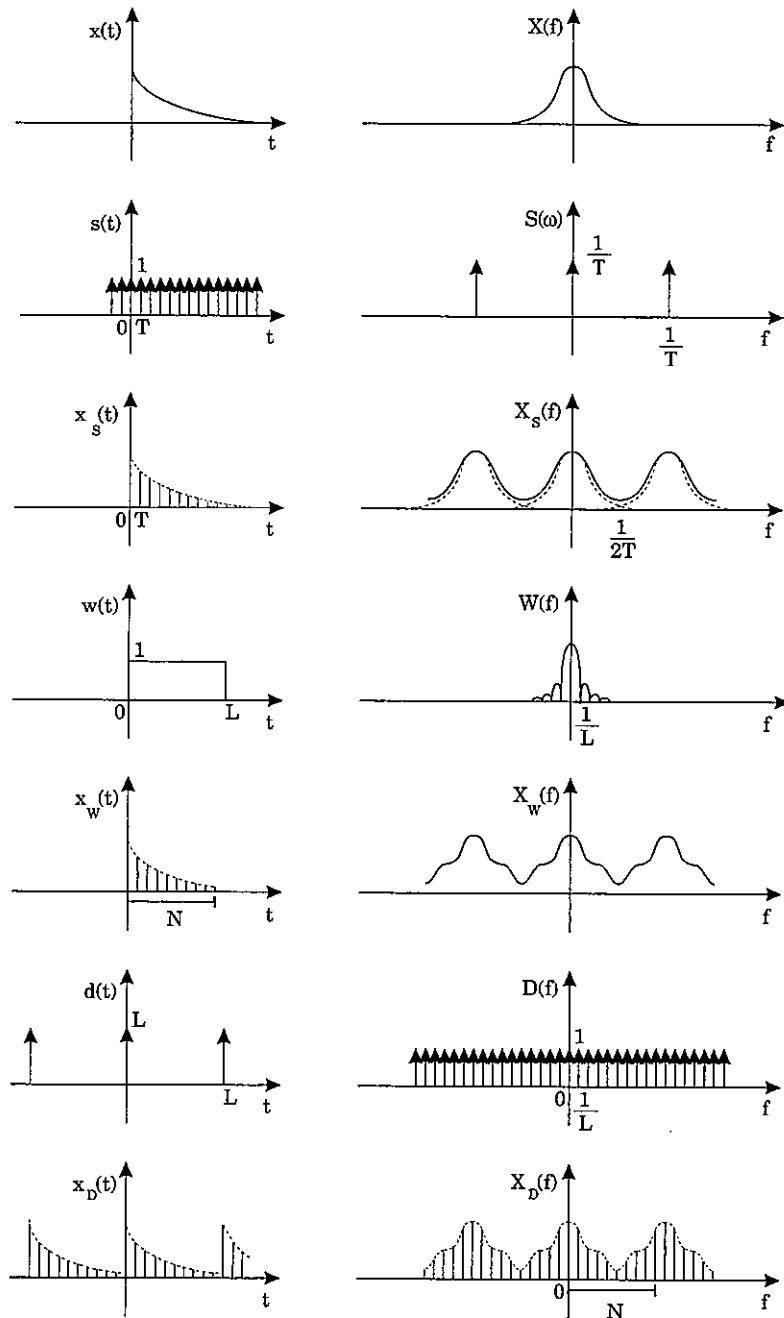



Fig. E.2 - Legame tra DTFT e DFT su di una sequenza di durata finita.

funzione stessa a multipli di un intervallo T (interpretabile come periodo di campionamento) (fig. E.2). Rappresentare una funzione tramite una serie (infinita) corrisponde a campionare la funzione stessa senza un preventivo filtraggio anti aliasing. Lo spettro della sequenza ottenuta dalla DTFT è quindi la ripetizione periodica dello spettro della funzione (eventualmente infinito) nell'intorno della frequenza F = 1/T, con un suo scalamento per la costante 1/T. A tale periodicizzazione è associato un fenomeno di aliasing. Passando alla DFT, considerare una serie finita di N elementi corrisponde a moltiplicare la serie precedentemente ottenuta dal campionamento con una rect(t) di ampiezza L = N*T. Dal punto di vista della rappresentazione in frequenza, tale troncamento corrisponde ad una convoluzione della DTFT con una sinc(f) (il cui primo zero è posto ad una frequenza pari a 1/L), che provoca una distorsione dello spettro tanto maggiore quanto più breve è la serie. La DFT di una serie di N elementi produce N campioni in frequenza ad intervalli F/N = 1/L. Tale campionamento in frequenza si riflette nella periodicizzazione nel tempo ad intervalli pari ad L. Fissato T, è possibile incrementare la definizione della rappresentazione in frequenza incrementando N tramite l'accodamento alla serie di campioni nulli (zero padding). Riducendo T, invece, si allontanano le repliche dello spettro, permettendo l'osservazione, di un maggiore range di frequenze (fig. E.3).

La diffusione della DFT è legata all'esistenza di una famiglia di algoritmi (Fast Fourier Transform: FFT) in grado di ridurre la complessità computazionale di una DFT da N^2 a $N \log_2 N$: infatti, definita la DFT di un segnale come

$$F(k) = \sum_{n=0}^{N-1} f(n)\, W_N^{nk} \qquad (E.11)$$

dove N è il numero di campioni, $W_N = e^{-j\frac{2\omega}{N}}$ la radice N-esima dell'unità, f(n) l'n-esimo campione del segnale e F(k) il k-esimo campione dello spettro del segnale, l'applicazione diretta della (E.11) comporta l'effettuazione di N prodotti tra i campioni e le potenze di W_N per ciascuno degli N campioni dello spettro, da cui la complessità di N^2 ricordata.

La FFT fornisce N campioni dello spettro del segnale in corrispondenza di frequenze, sia positive che negative, esprimibili come sottomultiple della frequenza di campionamento. Per segnali periodici, ciò vuol dire che campio-

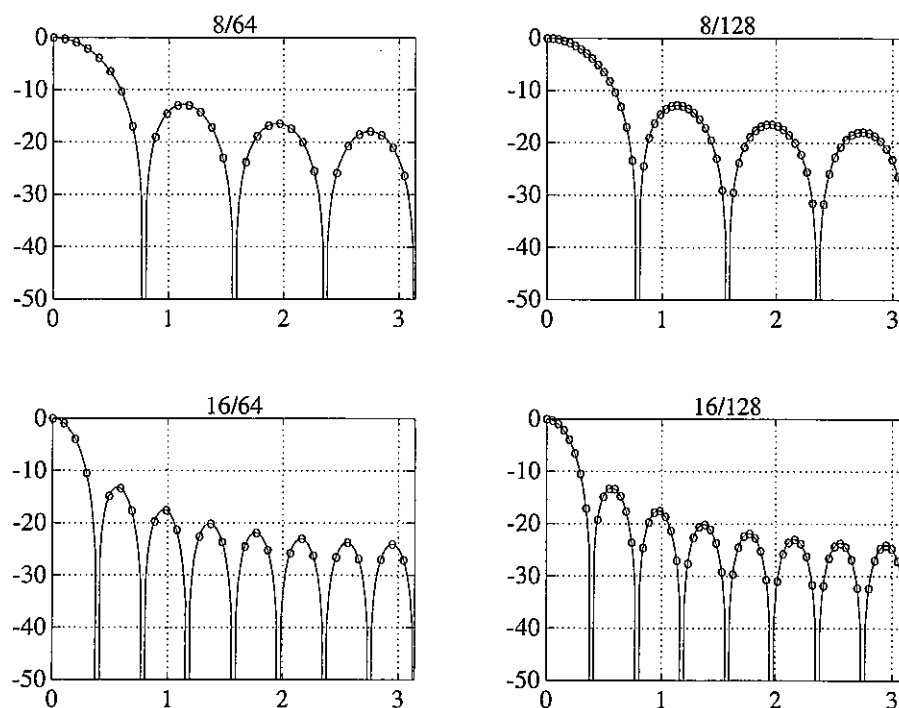



Fig. E.3 - DFT direct con numero di campioni e padding variabile.

nando N volte il segnale all'interno di un periodo, si avranno informazioni sulle prime N/2 armoniche.

Una prima classificazione di tali algoritmi può essere fatta in base alle radici dell'algoritmo stesso: si definiscono radici i fattori primi per mezzo dei quali è possibile scomporre il numero degli elementi della serie da trasformare. Ad esempio, gli algoritmi con radice due si applicano a blocchi di campioni in numero pari ad una potenza di due.

Una seconda classificazione degli algoritmi di FFT è tra algoritmi con decimazione nel tempo o in frequenza: tale distinzione si basa sul modo con il quale vengono organizzati i dati da elaborare. Con la decimazione nel tempo i campioni temporali del segnale da sottoporre a FFT vengono riordinati in modalità bit reversed nel dominio del tempo, mentre con la decimazione in frequenza la sequenza dei campioni non viene alterata, ma vanno riordinati i risultati della FFT.

L'idea base di tutti gli algoritmi di FFT è quella di scomporre una sequenza di campioni da trasformare in serie più brevi, le cui DFT, ricombinate, forniscano la DFT del segnale originario. In una FFT con radice pari a due, le serie da trasformare vengono ripetutamente dimezzate fino ad ottenere serie di due soli elementi, dopo $\log_2 N$ iterazioni. La DFT di una serie di due campioni risulta banalmente calcolabile, in quanto risulta essere dato da

$$\begin{cases} F(0) = f(0)\, W_2^{0*0} + f(1)\, W_2^{1*0} = f(0) + f(1) \\ F(1) = f(0)\, W_2^{0*1} + f(1)\, W_2^{1*1} = f(0) - f(1) \end{cases} \tag{E.12}$$

Si fa notare come per il calcolo di tale DFT non venga richiesto il calcolo di moltiplicazioni. Negli algoritmi di decimazione nel tempo la serie dei campioni è ripetutamente scomposta in due sottoserie, la prima delle quali contenente gli elementi di ordine pari della serie originaria, mentre la seconda quella di ordine dispari. L'ordinamento bit reserved deriva da tale progressiva scomposizione. È opportuno notare come le sottoserie generate possano essere viste come risultato di due campionamenti del segnale originario con una frequenza dimezzata ed istanti iniziali opportunamente sfasati. In tal modo la (E.11) può essere riscritta come

$$F(k) = \sum_{n=0}^{N/2-1} f(2n)\, W_N^{2nk} + \sum_{n=0}^{N/2-1} f(2n+1)\, W_N^{(2n+1)k}$$

$$F(k) = \sum_{n=0}^{N/2-1} f_1(n)\, W_{N/2}^{nk} + W_N^k \sum_{n=0}^{N/2-1} f_2(n)\, W_N^{2nk} \tag{E.13}$$

dove f_1 è la serie degli elementi pari, mentre f_2 è la serie degli elementi dispari. Il coefficiente W_N^k tiene, appunto, conto dello sfasamento tra le due sottosequenze.

Dato che $W_N^2 = W_{N/2}^1$ (fig. E.4) la (E.13) può essere scritta come

$$F(k) = \sum_{n=0}^{N/2-1} f_1(n)\, W_{N/2}^{nk} + W_N^k \sum_{n=0}^{N/2-1} f_2(n)\, W_{N/2}^{nk} \tag{E.14}$$

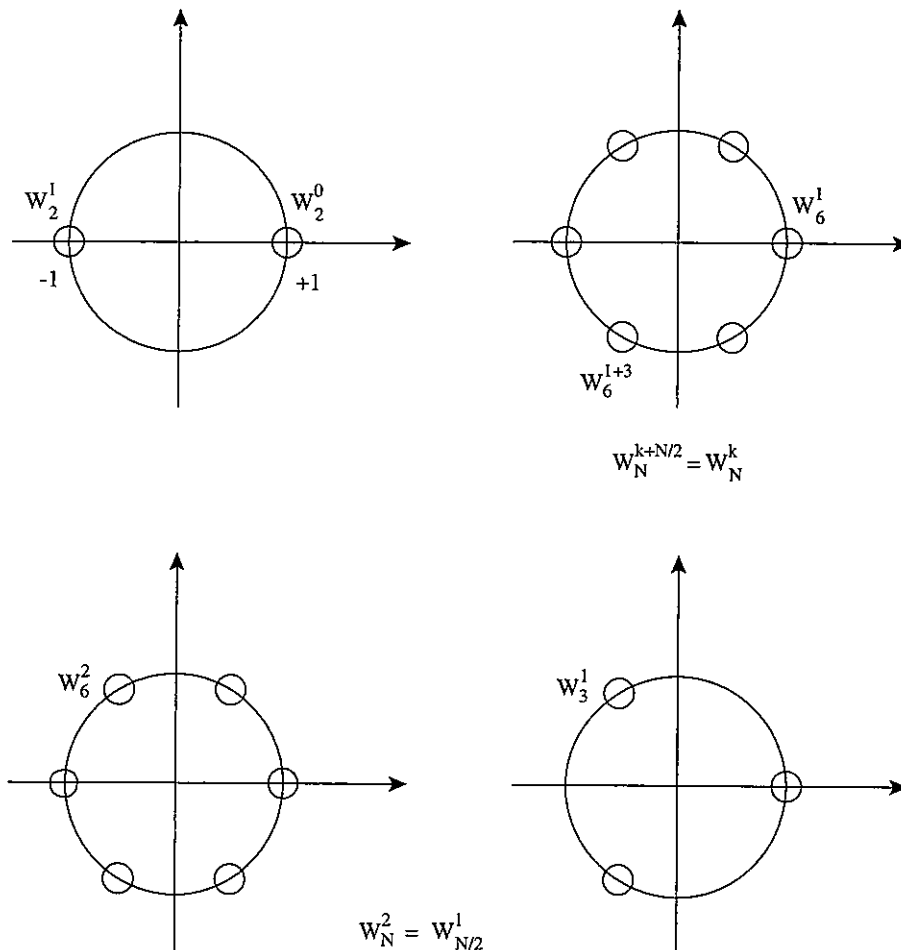



Fig. E4 - Proprietà della radice n-esima dell'unità

Applicando la (E.11), la (E.14) diventa

$$F(k) = F_1(k) + W_N^k F_2(k) \tag{E.15}$$

cioè, la DFT di una serie può essere ottenuta ricombinando secondo la (E.15) le trasformate di due sue opportune sottoserie ciò grazie alle proprietà di linearità e di traslazione nel tempo della trasformata di Fourier. Dato che la (E.15) deve valere per $0 \leq k \leq N-1$, essa va intesa come

$$F(k) = \begin{cases} F_1(k) + W_N^k F_2(k) & 0 \le k \le \frac{N}{2} - 1 \\ F_1\left(k - \frac{N}{2}\right) + W_N^k F_2\left(k - \frac{N}{2}\right) & \frac{N}{2} \le k \le N - 1 \end{cases} \quad \text{(E.16)}$$

Dato che $W_N^{k+N/2} = -W_N^k$, la (E.16) può essere riscritta come

$$F(k) = \begin{cases} F_1(k) + W_N^k F_2(k) & 0 \le k \le \frac{N}{2} - 1 \\ F_1\left(k - \frac{N}{2}\right) - W_N^{k-N/2} F_2\left(k - \frac{N}{2}\right) & \frac{N}{2} \le k \le N - 1 \end{cases} \quad \text{(E.17)}$$

In tal modo si vede, ad esempio, come la DFT di una serie di 8 campioni può essere ottenuta a partire dalle trasformate di due serie di quattro suoi elementi. Procedendo nella scomposizione di ciascuna sottoserie, si arriverà a poter applicare la (E.12). Lo spettro completo del segnale sarà ottenibile ricombinando le trasformate tramite ripetute applicazioni della (E.16).

Visto che la scomposizione del segnale in sottoserie si traduce nel riordinamento bit reversed dei campioni e che la valutazione della (E.12) non richiede l'effettuazione di nessuna moltiplicazione, l'ordine di complessità $N \log_2 N$ già ricordato deriva dall'applicazione della (E.16) a ciascuno dei $\log_2 N$ livelli di scomposizione ottenuti per ognuno degli N campioni dello spettro.

Inoltre, ora si può comprendere che il riordinamento bit reversed ha il compito di riordinare la serie dei campioni in modo da rendere adiacenti i campioni ai quali applicare la (E.12), che, in realtà, distano N/2 intervalli di campionamento; tali considerazioni valgono anche per le ulteriori sottosequenze in modo tale che la (E.16) possa essere applicata sequenzialmente, ottenendo in uscita la serie ordinata dei campioni dello spettro voluto (fig. E.5).

Come ultima considerazione si pone l'accento sul fatto che le (E.12) coincidono con le (E.14) scritte per N=2; ciò vuol dire che, ai fini del calcolo, non si distingue tra calcolo delle trasformate ricombinazione delle sottoserie. Il programma di calcolo della FFT si ridurrà all'applicazione iterativa della (E.16) ai campioni opportunamente ordinati, con N crescente dal valore 2 ad N/2 secondo le potenze di 2.

Es.: data la serie x(n) = {0, 2, 1, 1}, la sua FFT è data da

$$W_4^0 = 1;\ W_4^1 = -j$$

$$\begin{array}{lllll} x(00_b) & \rightarrow & x_0 + x_2 = 1 & \rightarrow & X(0) = 4 \\ x(10_b) & \rightarrow & x_0 - x_2 = -1 & \rightarrow & X(1) = -1 - j \\ x(01_b) & \rightarrow & x_1 + x_3 = 3 & \rightarrow & X(2) = 2 \\ x(11_b) & \rightarrow & x_1 - x_3 = 1 & > & X(3) = 1 + j \end{array} \tag{E.18}$$

E.2 DCT COME APPROSSIMAZIONE DELLA KLT

Si consideri un processo stazionario di Markov del primo ordine, per il quale la matrice di auto-covarianza è

$$[\,R\,]_{ik} = \rho^{|i-k|} \quad i, k = 0, 1, ..., N-1 \tag{E.19}$$

per $0 < \rho < 1$, dove ρ è il coefficiente di correlazione. La soluzione per la matrice Φ degli autovettori nel caso discreto è

$$[\,\Phi\,]_{mn} = \Phi_m(n) = \sqrt{\frac{2}{N+\mu_m}}\ \sin\left\{ w_n \left[(n+1) - \frac{(N+1)}{2} \right] + (m+1)\frac{\pi}{2} \right\}$$

$$m, n = 0, 1, ..., N-1 \tag{E.20}$$

dove $\mu_m = \dfrac{1-\rho^2}{1 - 2\cos(w_m) + \rho^2}$ sono gli autovalori e w_m sono le radici reali positive dell'equazione:

$$\tan(Nw) = -\frac{(1-\rho^2)\sin(w)}{\cos(w) - 2\rho + \rho^2\cos(w)} \tag{E.21}$$

Considerando il coefficiente di correlazione $\rho \rightarrow 1$, si ha $\tan(Nw) = 0$, da cui deriva

$$w_k = \frac{k\pi}{N}, \quad k = 0, 1, ..., N-1 \tag{E22}$$

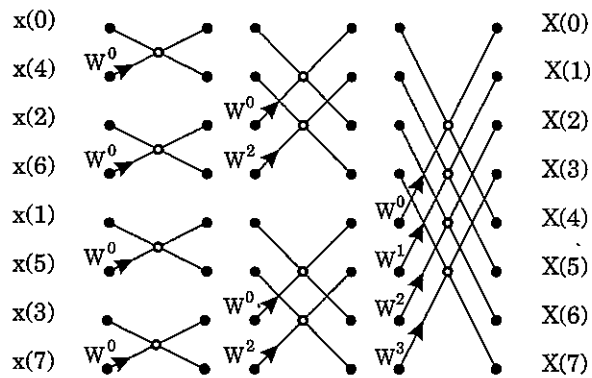
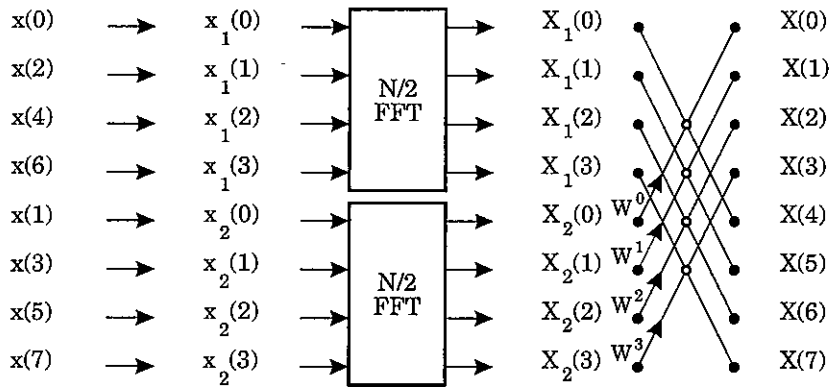
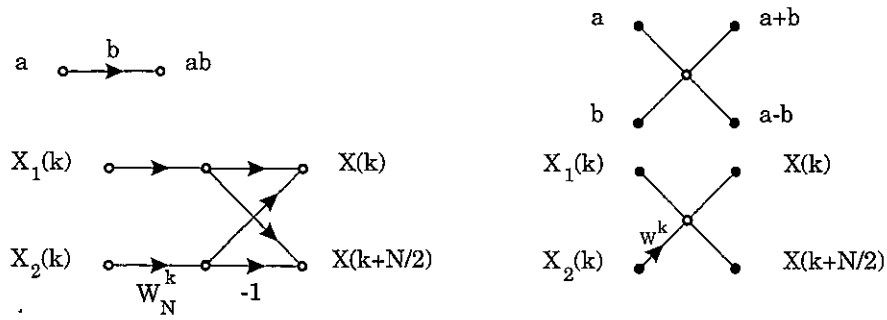



Fig. E.5 - Flusso dei dati in una FFT.

Gli autovalori μ_m sono uguali a zero quando i w_m sono diversi da zero. Per μ_m quando m = 0, l'espressione tende all'infinito. In conclusione, si ottiene che

$$\sum_{m=0}^{N-1} [A]_{mn} = \sum_{m=0}^{N-1} \mu_m \tag{E.23}$$

e, poiché gli elementi diagonali di [A] in un processo di Markov del primo ordine sono tutti unitari, abbiamo immediatamente che $\mu_0 = N$. Sostituendo si ottiene

$$\begin{cases} [\varnothing]_{n\,0} = \frac{1}{\sqrt{N}} \\ [\varnothing]_{n\,m} = \sqrt{\frac{2}{N}} \sin\left[m\left(n + \frac{1}{2} \right)\frac{\pi}{N} + \frac{\pi}{2} \right] \end{cases} \tag{E.24}$$

La seconda equazione può essere riscritta come

$$[\varnothing]_{n\,m} = \sqrt{\frac{2}{N}} \cos\left[m\left(n + \frac{1}{2} \right)\frac{\pi}{N} \right], \quad m \neq 0$$

$$= \sqrt{\frac{2}{N}}\, k_m \cos\left[m\left(n + \frac{1}{2} \right)\frac{\pi}{N} \right], \quad m, n = 0, 1, ..., N-1$$

$$k_m = \begin{cases} \frac{1}{\sqrt{2}} & m = 0 \\ 1 & \text{altrimenti} \end{cases} \tag{E.25}$$

Questa equazione è la definizione della DCT. Quindi la DCT è asintoticamente equivalente alla KLT per processi Markoviani del 1° ordine quando $\rho \to 1$. Si noti che ciò è indipendente da N e che l'unico autovalore diverso da zero è μ_0, pari ad N. In termini di MSE, l'errore è compattato in un singolo autovalore.

La ragione è che $\mathbf{R}_{-1}$ è una matrice simmetrica tri-diagonale, che per uno scalare $\beta_2 = (1 - \rho_2) / (1 + \rho_2)$ ed $\alpha = \rho/(1 + \rho^2)$ soddisfa la relazione

$$\beta^2 [A]^{-1} = \begin{bmatrix} 1-\rho\alpha & -\alpha & 0 & \dots \\ -\alpha & 1 & \dots & \dots \\ \dots & \dots & 1 & -\alpha \\ 0 & \dots & -\alpha & 1-\rho\alpha \end{bmatrix} \tag{E.26}$$

Si ricava l'approssimazione

$$\rho \approx 1 \;\rightarrow\; \beta^2 R^1 \approx \mathbf{Q}_y \tag{E.27}$$

Quindi gli autovettori di [A] e gli autovettori di $\mathbf{Q}_y$, cioè, la trasformata coseno, saranno molto simili.

BIBLIOGRAFIA

[Ata79] B.S. Atal, M.R. Schroeder «Predictive Coding of Speech Signals and Subjective Error Criteria» in *IEEE Transaction on ASSP*, vol. ASSP-27, n. 3, June 1979, pp. 247 - 254.

[Ata80] B.S. Atal, M.R. Schroeder «Improved quantizer for adaptive predictive coding of speech signals at low bit rates» Proc. of ICASSP, 1980, pp. 535 - 538.

[Ata82] B.S. Atal, J.R. Remde «A new model of LPC excitation for producing natural-sounding speech at low bit rates» Proc. of ICASSP, 1982, pp. 614 - 617.

[Ata84] B.S. Atal, M.R. Schroeder «Stochastic coding of speech signals at very low bit rates» Proc. of ICC, Amsterdam, 1984, pp. 1610-1613.

[Ata91] B.S. Atal, S. Bishnu, V. Cuperman, A. Gersho «Advances in Speech Coding» Kluwer Academic Publishers, 1991.

[Bat95] S. Battista, D. Sereno «MPEG-4: la futura codifica multimediale per applicazioni interattive» in *Alta Frequenza*, vol. 7, n. 5, Settembre-Ottobre 1995, pp. 43 - 49.

[Ber93] E. Berruto, D. Sereno «Variable-rate for the basic speech service in UMTS» Proc. of VTC, 1993.

[Bon91] L. Bonavoglia «Il Segnale Telefonico» SSGRR, 1991.

[Bra95] K. Brandenburg, M. Bosi «Overview of MPEG-Audio: Current and Future Standard for Low Bit-Rate Audio Coding» 99TH AES CONVENTION Preprint, 1995.

[Car94] F. Carassa «Divagazioni sui concetti di base della comunicazione» in *AEI*, vol. 81, n. 11, Novembre 1994, pp. 32-39.

[Cel89] L. Cellario, G. Ferraris, D. Sereno «A 2 ms delay CELP coder» Proc. of ICASSP, Glasgow, 1989, pp. 73-76.

[Cel94] L. Cellario, D. Sereno «CELP coding at variable rate» in *ETT*, vol. 5, n. 5, Settembre-Ottobre 1994, pp. 69-80.

[Cel96] L. Cellario, M. Festa, D. Sereno, J.M. Müller, B. Wächter «An object oriented generic audio coding architecture» Proc. of ICCT 96, Maggio 1996.

[Cop84] M. Copperi, D. Sereno «Improved LPC excitation based on pattern classification and perceptual criteria» Proc. of ICPR, Montreal, 1984, pp. 860-862.

[Cro83] R.E. Crochiere, L.R. Rabiner «Multirate Digital Signal Processing» Prentice-Hall, 1983.

[Dav86] G. Davidson, A. Gersho «Complexity reduction methods for vector excitation coding» Proc. of ICASSP, Tokyo 1986, pp. 3055 - 3058.

[Del93] J.R. jr Deller, J.G. Proakis, J.H.L. Hansen «Discrete-Time Processing of Speech Signals» Macmillan Publishing Company, 1993.

[Dro91] R. Drogo de Iacovo, D. Sereno «Embedded CELP coding for variable bit-rate between 6.4 and 9.6 kbit/s» Proc. of ICASSP, 1991, pp. 681 - 684.

[ETSI06.20] European Conference of Post and Telecommunications Administrations «Half-rate speech transcoding» ETSI GSM Recommendation 06.20, Gennaio 1995.

[ETSI06.32] «Voice activity detection» ETSI/GSM Recommandation 06.32.

[Fes95] M. Festa, D. Sereno, «A speech coding algorithm based on prototypes interpolation with critical bands and phase coding» Proc. of Eurospeech 1995, pp. 229 - 232.

[Fla72] Flanagan, L. James «Speech Analysis Synthesis and Perception» Springer-Verlag, 1972.

[Gel63] I.M. Gelfand, S.V. Fomin «Calculus of variations» Prentice-Hall, 1963.

[Goo90] D.J. Goodman «Cellular packet communications» in *IEEE Tr. on Communications*, vol. 38, Agosto 1990, pp. 1272-1280.

[Gra76] A. Gray, J.D. Markel, «Distance measures for speech processing» in *IEEE Tr. on ASSP*, vol. ASSP-24, n. 5, Ottobre 1976, pp. 380 - 391.

[GSM] European Conference of Post and Telecommunications Administrations «GSM full rate speech transcoding» CEPT/GSM Recommendation 06.10, 1989.

[Hay86] S. Haykin «Adaptive Filter Theory» Prentice-Hall, 1986.

[Hes83] W. Hess «Pitch determination of speech signals» Springer-Verlag, Berlin 1983.

[Hon84] M. Honda, F. Itakura «Bit allocation in time and frequency domains for predictive coding of speech» in *IEEE Tr. on ASSP*, vol. ASSP-32, n. 3, Giugno 1984, pp. 465 - 473.

[Ita75] F. Itakura «Line spectrum representation of linear predictive coefficients of speech signals» in *J. Acoust. Soc. Am.*, 57, 535(A), 1975.

[ITUT G.132] International Telecommunication Union «Attenuation Distortion» ITU-T Recommendation G.132, 1988.

[ITUT G.711] International Telecommunication Union «Pulse Code Modulation (PCM) of Voice Frequencies» ITU-T Recommendation G.711, 1972.

[ITUT G.712] International Telecommunication Union «Performance Characteristics of PCM Channels Between 4-wire Interfaces at Voice Frequencies» ITU-T Recommendation G.712, 1972.

[ITUT G.721] International Telecommunication Union «32 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)» ITU-T Recommendation G.721, 1984.

[ITUT G.722] International Telecommunication Union «7 kHz Audio-Coding within 64 kbit/s» ITU-T Recommendation G.722, 1988.

[ITUT G.727] International Telecommunication Union «General aspects of digital transmission systems; terminal equipments. 5,4,3 and 2 bits sample embed ded adaptive differential pulse code modulation (ADPCM)» ITU-T Recommendation G.727.

[ITUT G.802] International Telecommunication Union «Interworking between networks based on different digital hierarchies and speech encoding laws» ITU-T Recommendation G.802.

[ITUT P.50] International Telecommunication Union «Artificial Voices» ITU-T Recommendation P.50, 1988.

[ISO92] International Organization for Standardization/International Electrotechnical Commission «Information technology - Coding of moving pictures and associated audio for digital storage media up to about 1,5 Mbit/s» ISO/IEC JTC 1/SC 29/WG11 (MPEG) International Standard 11172, 1992.

[Jay76] N.S. Jayant «Waveforms Quantization and Coding» IEEE Press, 1976.

[Jay84] N.S. Jayant, P. Noll «Digital Coding of Waveforms - Principles and Applications to Speech and Video» Prentice-Hall, 1984.

[Kab86] P. Kabal, R.P. Ramachandran «The computation of line spectrum frequencies using Chebyshev polynomials» in *IEEE Tr. on ASSP*, vol. ASSP-34, n. 6, Dicembre 1986, pp. 1419-1426.

[Kay92] S. Kay «Extended-TDMA a high capacity evolution of US digital cellular» Proc. of Intern. Conf. Universal Personal Communications, Settembre 1992, pp. 07.04/1-3.

[Kin82] L.E. Kinsler, A.R. Frey, A.L. Coppens, J.V. Sanders, «Fundamentals of Acoustics» John Wiley & Sons, 1982.

[Kle91] W.B. Kleijn «Continuous representations in linear predictive coding» Proc. of ICASSP,1991, pp. 202 - 204.

[Kub93] G. Kubin, B.S. Atal, W.B. Kleijn «Performance of noise excitation for unvoiced speech» Proc. of IEEE workshop on speech coding for telecommunications, Ottobre 1993, pp. 35 - 36.

[Lin80] Y. Linde, A. Buzo, R. Gray «An algorithm for vector quantizer design» in *IEEE Transaction on Communications*, vol. 28, Gennaio 1980, pp.85-95.

[Lin88] D. Lin «Vector Excitation coding using a composite source model» Proc. of EUSIPCO,1988, pp. 859 - 862.

[Mak79] J. Makhoul, M. Berouti «Adaptive noise spectral shaping and entropy coding in predictive coding of speech» in *IEEE Tr. on ASSP*, vol. ASSp-27, n. 1, Febbraio 1979, pp. 63 - 73.

[Mar76] J.D. Markel, A.H. jr. Gray «Linear Prediction of Speech» Springer-Verlag, 1976.

[Moo89] C.J.B. Moore, «An Introduction to the Psychology of Hearing» Academic Press, 1989.

[Mus90] H.G. Musmann «The ISO Audio Coding Standard» *IEEE*, 1990.

[Pap84] A. Papouliss «Probability, Random Variables and Stochastic Processes» McGraw-Hill, 1984.

[Pap87] P.E. Papamichalis «Pratical Approaches to Speech Coding» Prentice-Hall, 1987.

[Pas94] E. Paskoy, K. Srinivasan, A. Gersho «Variable bit-rate CELP coding of speech with phonetic classification» in *ETT*, vol. 5, n. 5, Settembre-Ottobre 1994, pp. 57 - 68.

[Pro92] J.G. Proakis, D.G. Manolakis «Digital Signal Processing - Principles, Algorithms and Applications» Macmillan, 1992.

[Rab78] L.R. Rabiner, R.W. Shafer «Digital Processing of Speech Signals» Prentice-Hall, 1978.

[Sch85] M. R. Schroeder, B.S. Atal «Code-Excited Linear Prediction (CELP): high quality speech at very low bit rates» Proc. of ICASSP, Tampa 1985, pp. 937-940.

[Sha71] C.E. Shannon, W. Weaver «La teoria matematica delle comunicazioni» Etas Kompass, 1971.

[Sho93] Y. Shoham «High-quality speech coding at 2.4 to 4.0 kbps based on time-frequency interpolation» Proc. of ICASSP, 1993, pp. II-167/170.

[Soo84] F.K. Soong, B.H. Juang «Line spectrum pair (LSP) and speech data compression» Proc. of ICASSP 84, paper 1.10, 1984.

[Tan94] Y. Tanaka, H. Kimura «Low-bit-rate speech coding using a two dimensional transform of residual signals and waveform interpolation» Proc. of ICASSP 1994, pp. I-173/176.

[Tau77] H. Taub, D. Schilling «Elettronica Integrata Digitale» McGraw-Hill, 1977.

[Toh79] Y. Tohkura «Experimental comparison of parameter interpolation in LPC vocoders» Proc. of 97th MEETING OF ACOUSTICAL SOCIETY OF AMERICA, June 1979, pp. 373-376.

[Tsu92] T. Kyoya, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, R.M. Heddle «ATRAC: Adaptive Transform Acoustic Coding for MiniDisc» 93rd AES CONVENTION Preprint, 1992.

[Un75] C.K. Un, D.T. Magill «The Residual-Excited linear prediction Vocoder with transmission rate below 9.6 kbit/s» in *IEEE Tr. on Comm.*, vol. com-23, n. 12, Dicembre 1975, pp. 1466-1474.

[Wak81] H. Wakita «Linear prediction voice synthesizers: Line spectrum pairs (LSP) is the newest of several techniques» in *Speech Technol.*, Fall 1981.

[Wir91] G.C. Wirtz «Digital Compact Cassette: The Audio Coding Technique» 91th AES CONVENTION Preprint, 1995.